Partial NIR-VIS Heterogeneous Face Recognition With Automatic Saliency Search

Mandi Luo, *Member, IEEE*, Xin Ma, Zhihang Li, Jie Cao, *Member, IEEE*, and Ran He, *Senior Member, IEEE*

***Abstract*— Near-infrared-visual (NIR-VIS) heterogeneous face recognition (HFR) aims to match NIR face images with the corresponding VIS ones. It is a challenging task due to the sensing gaps among different modalities. Occlusions in the input face images make the task extremely complex. To tackle these problems, we present a Saliency Search Network (SSN) to extract domain-invariant identity features. We propose to automatically search the efficient parts of face images in a modality-aware manner, and remove redundant information. Moreover, the searching process is guided by an information bottleneck network, which mitigates the overfitting problems caused by small datasets. Extensive experiments on both complete and partial NIR-VIS HFR on multiple datasets demonstrate the effectiveness and robustness of the proposed method to modality discrepancy and occlusions.**

***Index Terms*— Heterogeneous face recognition, near infrared-visible matching, information bottleneck, neural architecture search.**

I. Introduction

FACE recognition in controlled environments has achieved stunning results [1]–[5], with accuracy even higher than that achieved by human beings. However, the performance of face recognition methods in real circumstances is still restricted by some bottleneck factors, including variations of sensing modalities, illumination, and so on [6]–[12]. These factors cause severe domain discrepancies that result in performance degradation, thus prompting the task of Heterogeneous Face Recognition (HFR). In this paper, we focus on near-infrared-visual (NIR-VIS) HFR [13], which aims to match captured partial near-infrared (NIR) images with the corresponding visual (VIS) images. Specifically, NIR cameras provide an inexpensive solution to capture face images in extreme lighting conditions while preserving identity information to the utmost extent. Thus, they are widely assembled in mobile devices, monitoring video cameras, and other applications. Since the enrolled template face images are usually in the VIS domain, NIR-VIS HFR is urgently needed.

NIR-VIS HFR has been widely used in security fields, including e-commerce, and security checks. Recently, under this COVID-19 pandemic period, to facilitate epidemiological investigations and epidemic prevention, personnel control has been strengthened, where NIR-VIS HFR plays an essential role. In traditional NIR-VIS HFR, complete NIR face images are needed during the matching process. However, it has become a new normal that people go out wearing their goggles and/or masks to avoid COVID-19. The risk of infection will increase if people take off their masks each time complete NIR facial images must be obtained. Thus, matching occluded NIR face images with enrolled VIS images, termed as partial NIR-VIS HFR, is desiderata.

Some studies have focused on NIR-VIS HFR. Recently, methods based on deep learning networks have attracted considerable attention. For instance, Liu *et al.* [14] proposed a triplet loss to reduce intra-class variations, as well as to augment the training dataset. Saxena and Verbeek [15] utilized a CNN pre-trained on VIS images to perform HFR. Wu *et al.* [16] introduced the Disentangled Variational Representation (DVR) to decouple an intrinsic variable for identity in both the NIR and VIS face images. These methods attempt to explore domain-invariant features in face images of different modalities; in other words, they only focus on identity-related information in face images and ignore other features.

However, while dealing with partial NIR-VIS HFR, these domain-invariant methods still face challenges that are three-fold: 1) Domain discrepancies. Different sensory devices are adopted to capture NIR and VIS face images of the same subject, leading to big appearance differences. Moreover, NIR face images are often captured at low resolutions under extreme lighting conditions. These images lose some identity-related

Manuscript received May 9, 2021; revised August 13, 2021 and September 29, 2021; accepted October 6, 2021. Date of publication October 21, 2021; date of current version November 5, 2021. This work was supported in part by the Beijing Natural Science Foundation under Grant JQ18017, in part by the National Natural Science Foundation of China under Grant U20A20223 and Grant 61721004, and in part by the Youth Innovation Promotion Association Chinese Academy of Sciences (CAS) under Grant Y201929. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William R. Schwartz. *(Corresponding author: Ran He.)*

Mandi Luo, Xin Ma, and Ran He are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Center for Research on Intelligent Perception and Computing, Chinese Academy of Sciences, Beijing 100864, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: luomandi2019@ia.ac.cn; xin.ma@cripac.ia.ac.cn; rhe@nlpr.ia.ac.cn).

Zhihang Li and Jie Cao are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: zhihang.li@nlpr.ia.ac.cn; jie.cao@cripac.ia.ac.cn).

Digital Object Identifier 10.1109/TIFS.2021.3122072

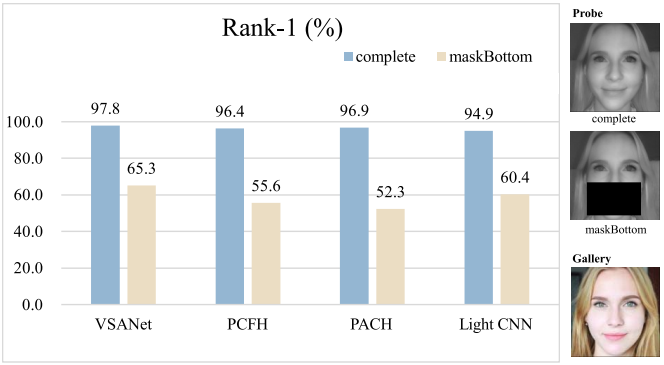


Fig. 1. Rank-1 face recognition rates (%) of VSANet [17], PCFH [18], PACH [19], and Light CNN [20] on the LAMP-HQ dataset. The labels "complete" and "maskBottom" indicate that the corresponding probes are complete and occluded at the lower half of the face, respectively, as the face images shown on the right side.

texture information, such as part of the hair, face cheek, and so on. Thus, it is quite challenging to meet the inter-modality gap and match NIR face images with the VIS ones. 2) Occlusion. Admittedly, face organs play an important part in NIR-VIS HFR. Extra occlusions on the input NIR face will introduce intra-modality discrepancies between partial NIR face images with complete ones, severely disrupting the model performance. As shown in Fig. 1, if we occlude the lower half of the face, the recognition accuracy of existing methods drops dramatically by 33%∼46%. 3) Small dataset. The scale of NIR-VIS datasets [17], [21]–[23] is rather small compared with that of other face datasets. For instance, the largest and newest dataset in NIR-VIS HFR, namely LAMP-HQ [17], contains 573 subjects with 73,617 images in total. There are approximately 10M images of 100k celebrities in the training set of the commonly used VIS face dataset MS-Celeb-1M [24], which is notably larger than the LAMP-HQ dataset. Training models with small datasets is more likely to suffer from the overfitting problem.

We aim at developing a framework to tackle the aforementioned issues. Luckily, We found that the recognition models focus on different face parts while dealing with images from different modalities or different subjects of the same modality. As shown in Fig. 2, we present the visualization of feature maps produced by Light CNN [20]. The more red a map area is, the more active the corresponding pixels in the recognition process. Thus, if we develop a mechanism that can automatically and self-adaptively find the most efficient and active parts of every face image, namely, the salient field, we can reduce the interference from inactive parts, thus decreasing both inter- and intra-modality discrepancies.

Based on this assumption, we propose a Saliency Search Network (SSN) with a novel pixel selection block (PSB) responsible for searching salient fields at the pixel level. Images from different domains own different PSBs. Note that every pixel is selected with a specific weight between 0 and 1. This strategy enables the active pixels to participate more in face recognition. Either the inactive parts of complete NIR face images or the occluded pixels in partial NIR face images are disabled, thus enhancing the quality and efficiency for representation learning.

Note that it is not easy for models to search salient fields with manual design. The randomness of salient fields is reflected in position and intensity. In other words, every pixel can be selected or abandoned, with the weights of chosen pixels range from 0 to 1, leading to a large number of combinations that form an enormous search space. Moreover, the salient field of every face image differs from each other, making the process extremely challenging. Inspired by the Neural Architecture Search (NAS) strategy, we introduce an automatic feature search (AFS) algorithm to perform the search process to improve the network's efficiency and accuracy. The AFS algorithm automatically and self-adaptively adjusts the selected salient fields based on the validation results. Specifically, we follow DARTS [25] to perform a continuous search that is compatible with the stochastic gradient search technique.

To address the overfitting and false correlation problems caused by small datasets, we further equip our saliency search with an information bottleneck (IB) trade-off. The information bottleneck aims to compress inputs without sacrificing the ability to accurately predict the labels. Only preserves identity-relevant information is preserved, and other information is compressed out. Thus, the proposed SSN is able to produce optimal feature representations.

We conduct extensive experiments on multiple datasets, including the CASIA NIR-VIS 2.0 Face dataset [22], the Oulu-CASIA NIR-VIS dataset [23], the BUAA-VisNir Face dataset [21], and the LAMP-HQ dataset [17]. Qualitative and quantitative analyses demonstrate that our proposed network surpasses other methods in the task of NIR-VIS HFR, especially partial NIR-VIS HFR.

Our main contributions can be summarized as follows,

- Based on our findings that the effective parts of face images in different modalities differ from each other, we propose a novel modality-aware network, namely Saliency Search Network (SSN), to explore domain-invariant features in the task of NIR-VIS HFR. The proposed pixel selection block (PSB) enables the active parts for face recognition, as well as disable inactive parts. The selection strategy reduces interference from redundant information.
- We present an automatic feature search (AFS) algorithm, which automatically optimizes the searching results to produce the optimal solution. The searching efficiency and accuracy are drastically improved.
- We introduce an information bottleneck (IB) as guidance for the search process. By adjusting the IB trader-off, our proposed SSN is able to address the overfitting and false correlation problems caused by small-scale datasets.
- Extensive qualitative and quantitative results prove that the proposed SSN performs better than other existing methods in the task of NIR-VIS HFR, especially when extra occlusions are introduced.

II. Related Work

A. Face Recognition

Face recognition has always been an active area for years since it plays an important role in both academic research and

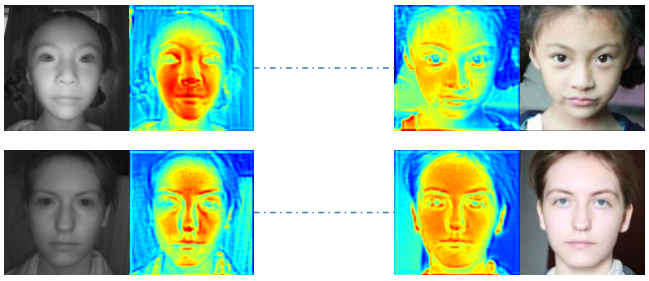


Fig. 2. Visualization of feature maps produced by Light CNN [20]. The dashed line indicates that the images are from the same identity.

real-world applications. Researchers have made tremendous efforts to push the frontiers [1]–[12], [26]–[28]. For instance, Zhao *et al.* paid attention to pose-invariant face recognition and propose a series of works. They designed advanced Generative Adversarial Networks (GANs) with dual paths [6] or dual agents [5], and introduced the information from 3D faces to 2D ones [7], [26] to better recover the lost information. They achieved photorealistic and identity preserving profile/frontal face synthesis even under extreme poses. Apart from poses, there are also many other factors influencing the face recognition accuracy, including illumination, age, and so on [9], [11]. All these challenges inspire researchers to keep moving forward and exploring.

1) Heterogeneous Face Recognition: Heterogeneous face recognition (HFR) [29], [30] refers to face recognition and matching across different visual domains. In most cases, the gallery for heterogeneous face recognition problems consists of visible light photos, while probes are pictures from other visual domains. According to the different modalities of the face, the current research on heterogeneous face recognition [31] mainly includes Sketch-VIS, NIR-VIS, 3D-2D, High-Low Resolution, Profiles-Frontal Face, and so on.

The existing approaches for NIR-VIS HFR can be mainly divided into three categories, including image synthesis methods, latent subspace methods, and domain-invariant feature methods. The intuitive idea of the image synthesis methods [17], [31] is to transform an image of a certain modality into an image of another modality, and then perform matching and recognition tasks in the same modality. The synthesized images are directly matched with an existing traditional face recognition model, so its performance and robustness mainly depend on the performance of the picture synthesis method. Latent subspace methods [32], [33] are to project two different modal images into the same subspace, so that they can be better compared. Domain-invariant features methods [34], [35] focus on obtaining facial features that are consistent among different modalities.

In this paper, we focus on domain-invariant feature methods. For example, Liao *et al.* [36] utilized Difference-of-Gaussian filtering to obtain a normalized appearance for all heterogeneous faces. They then applied MB-LBP to encode local image structures and learned the most discriminant local features. Shao and Fu [37] proposed a novel hierarchical hyperlingual-words (Hwords) framework, as well as a weighted distance metric, to help address to help perform HFR for images with pose and expression variations. Saxena and Verbeek [15] first attempted to use CNNs pretrained on visible spectrum images to perform HFR tasks and achieved satisfying results. This strategy was further adopted by many other researchers. Notably, Liu *et al.* [14] employed an activation function, namely, Max-Feature-Map, to select discriminative features. Reale *et al.* [38] developed a way that used large-scale visible face recognition datasets to learn global features in the task of HFR. Sarfraz and Stiefelhagen [35] captured the highly nonlinear relationships among different modalities. He *et al.* [34] aimed to minimize the Wasserstein distance among the distributions of different modalities with the proposed Wasserstein CNN. Du *et al.* [39] paid attention to NIR-VIS masked face recognition and addressed the problem considering both the training data and the training method. They adopted a 3D face reconstruction approach to synthesize masked faces and proposed an HSST method to extract domain-invariant face feature representations with a semi-Siamese network. Our proposed SSN also aims at extracting domain-invariant features, however, our process is automatic and self-adaptive.

2) Partial Face Recognition: Occlusions on faces pose a major challenge in the task of face recognition [40], [41]. Researchers have achieved important results from two main perspectives. Some of them [42], [43] attempted to generate clean and complete faces from the occluded ones; others [44], [45] focused on extracting local face representations that only from the nonoccluded areas. For instance, Min *et al.* [46] proposed using Gabor wavelets, PCA, and Support Vector Machines (SVM) to address the influence of scarfs/sunglasses in face recognition. Park *et al.* [42] introduced a Scale Invariant Feature Transform (SIFT) method to measure the similarity and match occluded images with the nonoccluded images. Song *et al.* [47] proposed using a Pairwise Differential Siamese Network (PDSN) to exploit the differences between occluded and nonoccluded faces; in their approach, a Feature Discarding Mask (FDA) is generated accordingly to express the correspondence between occluded facial areas and corrupted feature elements. Note that unlike these aforementioned methods that only pay attention to occlusions in the visual domain, we aim at addressing the occlusions across different domains, which is much more challenging.

B. Neural Architecture Search

Deep learning can be employed to automatically learn useful features, breaking away from the dependence on feature engineering, and achieve excellent results that surpass those of other algorithms in many tasks, including generation, segmentation, recognition, and so on. This success is largely due to the emergence of neural network structures, such as ResNet [48], DenseNet [49], etc. However, designing a high-performance neural network requires abundant professional knowledge and extensive trial, and the cost is extremely high; these factors limit the development and application of neural networks. Neural Architecture Search (NAS) [25], [50] is a technology

for automatically designing neural networks. It automatically designs high-performance network structures based on sample sets through algorithms. The performance of this approach even matches the level achieved by human experts in certain tasks, and some network structures that have not been proposed by humans before have been discovered; consequently, the use and implementation costs of neural networks have been effectively reduced. NAS technology has been widely applied in many areas, including object detection [51], semantic segmentation [52], image classification [53], and other fields. Inspired by NAS, we propose an automatic feature search algorithm to help extract the active face parts in NIR-VIS HFR.

C. Information Bottleneck

The information bottleneck (IB) concept was originally introduced by Tishby *et al.* [54] as an information theory. Given the joint probability $\mathbb{P}(\mathbf{X}, \mathbf{Y})$ between a random variable $\mathbf{X}$ and an observed relevant variable $\mathbf{Y}$, IB aims at extracting the best trade-off between complexity and accuracy while clustering $\mathbf{X}$. In 2017, Alemi *et al.* [55] first introduced this theory to the deep learning field and proposed a variational approximation to IB. With the development of deep learning, the theory of IBs has been widely used in many areas [56], [57], including natural language processing (NLP) [58], reinforcement learning [59], [60], graph encoding [61], and so on. IBs have also received considerable attention in the computer vision field, which is this focus of this study. For example, Peng *et al.* [62] proposed to using an information bottleneck as a regularization to constrain information flow in the discriminator. The introduced variational discriminator bottleneck (VDB) effectively improved the performance of Generative Adversarial Networks (GANs) in the task of image generation. Luo *et al.* [63] introduced a significance-aware information bottleneck (SIB) to ease the feature alignment and stabilize adversarial training in the task of unsupervised semantic segmentation. Similar to these aforementioned methods, our method uses the IB concept as a trade-off mechanism in the proposed SSN to compress input images while achieving the highest predictive power possible.

III. Methods

We adopt the Light CNN [20], which has been widely employed in the task of NIR-VIS HFR, as the baseline in the proposed Saliency Search Network (SSN). Given an input NIR face image $\mathbf{I}_{in}$, Light CNN encodes $\mathbf{I}_{in}$ into an identity embedding z which is as close as possible to images in the corresponding gallery VIS face images $\mathbf{I}_{gt}$, i.e., z_{gt}. During the training phase, this optimization process is supervised by a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{id} &= -\sum_{k} z_{gt}^{k} \log\left(z^{k}\right) \\ &= -\sum_{k} L(\mathbf{I}_{gt}^{k}) \log\left(L(\mathbf{I}_{in}^{k})\right), \end{aligned} \tag{1}$$

where k denotes the identity index for a specific person and L(·) denotes the discriminative features extracted by Light CNN.

As shown in Fig. 3, the gray boxes denote the original architecture of Light CNN. Note that Light CNN is composed of several similar blocks, where each contains a residual block and a Max-Feature- Map (MFM) group. An MFM [20] is a special implementation of maxout activation [64] to suppresses low-activation neurons in each layer. For each residual block B, given an input feature map $\mathbf{X}_B \in \mathbb{R}^{C_{in}\times W\times H}$, the output feature map is denoted as $\mathbf{Y}_B \in \mathbb{R}^{C_{\text{out}}\times W'\times H'}$, where C, W, and H denote the number of channels, width, and height, respectively. For a specific position $p \in V$ ($V = \{(i, j) \mid i \le W, j \le H, i, j \in \mathbb{Z}^{+}\}$), given the support region R_r with size r, the corresponding $\mathbf{Y}_B(p)$ is computed as:

$$\mathbf{Y}_B(p) = \sum_{p'\in R_r} \theta_c\left(p'\right)\mathbf{X}\left(p+p'\right) \quad p \in V, \tag{2}$$

where θ_c indicates the convolutional kernel weights.

To improve the efficiency and accuracy of Light CNN, we propose to search salient fields, i.e., a subset $\mathbf{y}$ of every $\mathbf{Y}_B$. The subset preserves identity-related information as much as possible, and redundant information is neglected. In other words, given output feature maps $\mathbf{Y}_B$, we detect salient fields $\mathbf{y} \in \mathbf{Y}_B$, which can minimize the network loss $\mathcal{L}(\mathbf{y}, \theta)$ after minimizing the network weights θ. The objective function is thus formulated as:

$$\min_{\mathbf{y}\in\mathbf{Y}_B} \min_{\theta} \mathcal{L}(\mathbf{y}, \theta). \tag{3}$$

To automatically and self-adaptively search salient fields, our SSN is equipped with modality-aware pixel selection blocks (PSBs) and an information bottleneck-guided search algorithm to define the search space and perform the saliency search, respectively. The details of every component are discussed in the following parts. We use the presence and absence of the superscript $\hat{}$ on a variable to indicate whether it is drawn from the distribution of the input data or that of the output data, respectively.

A. Modality-Aware Pixel Selection Block

Different from common NAS methods that aim to search for the optimal topology of some given network architectures, we focus on searching on the space formed by the extracted features. Specifically, we propose the pixel selection block (PSB) to perform a pixel-level search in every intermediate feature map $\mathbf{Y}_B$ of the residual block B. As discussed in Sec. I, the active parts of a given face image change with both identity and modality variations. Thus, as shown in Fig. 3, we apply different PSBs according to the different modalities of the input face images. $\mathbf{Y}_B$ is then further decomposed into $\mathbf{Y}_B^{nir}$ and $\mathbf{Y}_B^{vis}$, which are feature maps extracted from images of different illumination modalities. The complete search space is then composed of all $\mathbf{Y}_B^m$, where $B \in \{1, \ldots, N\}$ and $m \in \{nir, vis\}$. Note N is the number of residual blocks.

Given a feature map $\mathbf{Y}_B^m \in \mathbb{R}^{C_{\text{out}}\times W'\times H'}$, a corresponding pixel-level saliency indicator $\mathbf{M}_B^m \in \mathbb{R}^{C_{\text{out}}\times W'\times H'}$ is initialized with a value of 1 at every pixel position, indicating that each pixel in the spatial domain is selected 100% at the beginning of the search process. During optimization, $\mathbf{M}_B^m$ is activated by

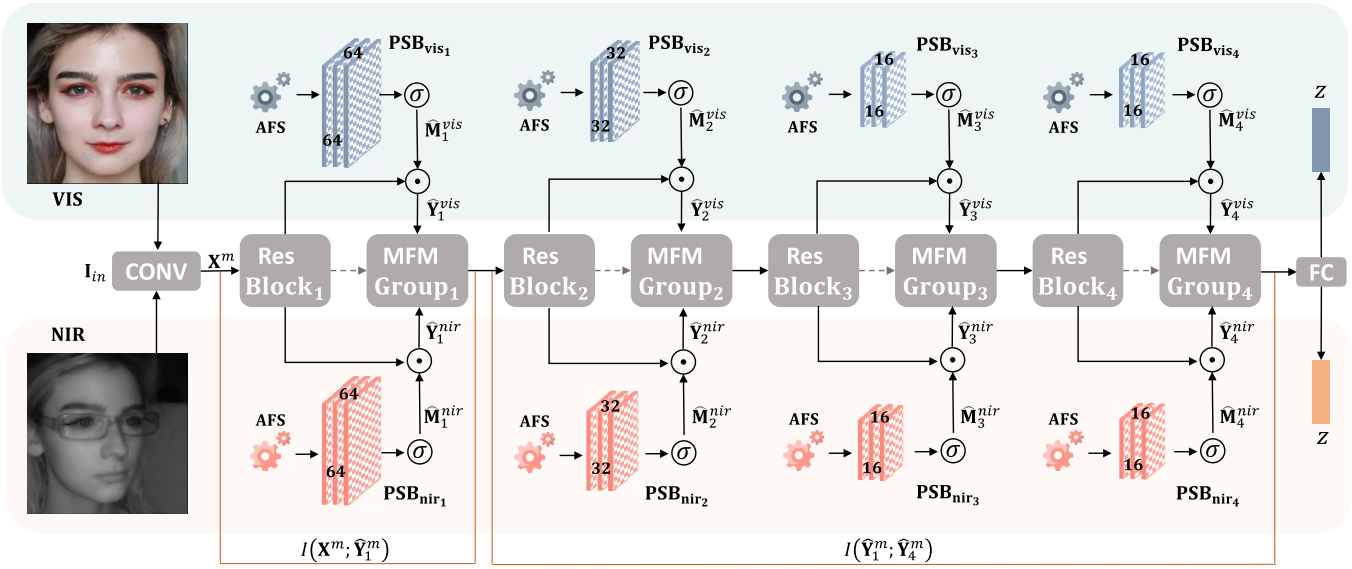


Fig. 3. Overall architecture of the proposed Saliency Search Network (SSN). Our network is developed based on Light CNN [20], and the original Light CNN architecture is represented in gray boxes. The gray dashed arrows indicate the original data flows. Given an input image $\mathbf{I}_{in}$, we propose a modality-aware search strategy to preserve the identity-related pixels of encoded feature maps, as well as vanish the redundant ones. The selection process is implemented by the pixel selection block (PSB). Note that Light CNN is composed of several similar blocks, where each of them contains a ResBlock and an MFM Group. Specifically, for each Light CNN block, we apply modality-aware PSBs. We first conduct an automatic feature search (AFS) to get the saliency indicator $\mathbf{M}_B^m$. Then, we obtain the selected feature map $\hat{\mathbf{Y}}_B^m = \sigma(\mathbf{M}_B^m) \odot \mathbf{Y}_B^m$, where σ, m, and B denote the sigmoid function, the modality of the input image, and the number of Light CNN blocks, respectively. Note that the search process is guided by an information bottleneck module, where the mutual information $I(\mathbf{X}_m; \hat{\mathbf{Y}}_m^1)$ is maximized, and $I(\mathbf{Y}_m^1; \hat{\mathbf{Y}}_m^4)$ is minimized.

a sigmoid function σ. Thus, the indicator is restricted between 0 and 1, indicating the weight that corresponding to a specific position p in the extraction of identity information. Then, we get parameterized output activation map $\hat{\mathbf{Y}}_B^m$ as:

$$\hat{\mathbf{Y}}_B^m = \sigma(\mathbf{M}_B^m) \odot \mathbf{Y}_B^m, \quad B \in \{1, \ldots, N\}, \ m \in \{nir, vis\}, \quad (4)$$

where $\odot$ denotes the Hadamard product.

B. Information Bottleneck-Guided Automatic Feature Search

1) Information Bottleneck: Note that the existing datasets used for NIR-VIS HFR are often in small-scale, resulting in overfitting and false correlation problems. To tackle these issues and extract the most expressive information, we introduce an information bottleneck trade-off to guide the search process. Specifically, given the input $\mathbf{X}^m$, the goal of our information bottleneck (IB) is to find an optimal representation that: 1) captures identity-related information as much as possible and 2) compresses the identity-irrelevant parts of the input $\mathbf{X}^m$ to the greatest extent. Here, we select the outputs of the first and last residual blocks as the intermediate and final representations, i.e., $\hat{\mathbf{Y}}_1^m$ and $\hat{\mathbf{Y}}_4^m$, respectively. The optimization of IB is then denoted as:

$$\mathcal{L}_{ib} = \min_{\hat{\mathbf{Y}}_1^m} I(\mathbf{X}^m; \hat{\mathbf{Y}}_1^m) - \beta I(\hat{\mathbf{Y}}_1^m; \hat{\mathbf{Y}}_4^m), \quad (5)$$

where β denotes the positive Lagrange multiplier that acts as a trade-off parameter. $I(\mathbf{X}^m; \hat{\mathbf{Y}}_1^m)$ defines the mutual information that represents the relevance of $\mathbf{X}^m$ and $\hat{\mathbf{Y}}_1^m$. The smaller the $I(\mathbf{X}^m; \hat{\mathbf{Y}}_1^m)$, the less relevant $\hat{\mathbf{Y}}_1^m$ is to $\mathbf{X}^m$. Thus, after optimizing Eqn. 5, $\hat{\mathbf{Y}}_1^m$ is as less relevant as possible to the input $\mathbf{X}^m$, while as more relevant as possible to the final identity representation $\hat{\mathbf{Y}}_4^m$. In this way, the intermediate representation $\hat{\mathbf{Y}}_1^m$ acts as an information bottleneck where only the most identity-relevant information is preserved.

Algorithm 1 AFS: Automatic Feature Search

Input: the saliency indicator $\mathbf{M}^m$, the network weights θ'', the training set $\mathbb{S}_T$, and the testing set $\mathbb{S}_E$
Output: the trained network and the optimal feature
1. Split $\mathbb{S}_T$ into $\mathbb{S}_{train}$ and $\mathbb{S}_{val}$, which responsible for training and valiadation during the training phase, respectively. Note each of them takes 50% of $\mathbb{S}_T$
2. **while** *not converged* **do**
  Update the saliency indicator $\mathbf{M}_B^m$ by descending $\nabla_{\mathbf{M}_B^m} \mathcal{L}_{\text{val}}(\theta'', \mathbf{M}_B^m)$
  Update the network weight θ'' by descending $\nabla_{\theta''} \mathcal{L}_{\text{train}}(\theta'', \mathbf{M}_B^m)$
end
3. Train the network weight θ'' with the searched $\mathbf{M}_B^m$ on $\mathbb{S}_T$
4. Evaluate on $\mathbb{S}_E$.

As declared in [65], the mutual information is defined by the Kullback-Leibler (KL-) divergence:

$$I(\mathbf{X}^m; \hat{\mathbf{Y}}_1^m) = D_{KL}\left(\mathbb{P}_{\mathbf{X}^m \hat{\mathbf{Y}}_1^m} \| \mathbb{P}_{\mathbf{X}^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_1^m}\right), \quad (6)$$

where D_{KL} is defined as,

$$D_{KL}(\mathbb{P} \| \mathbb{Q}) := \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right], \quad (7)$$

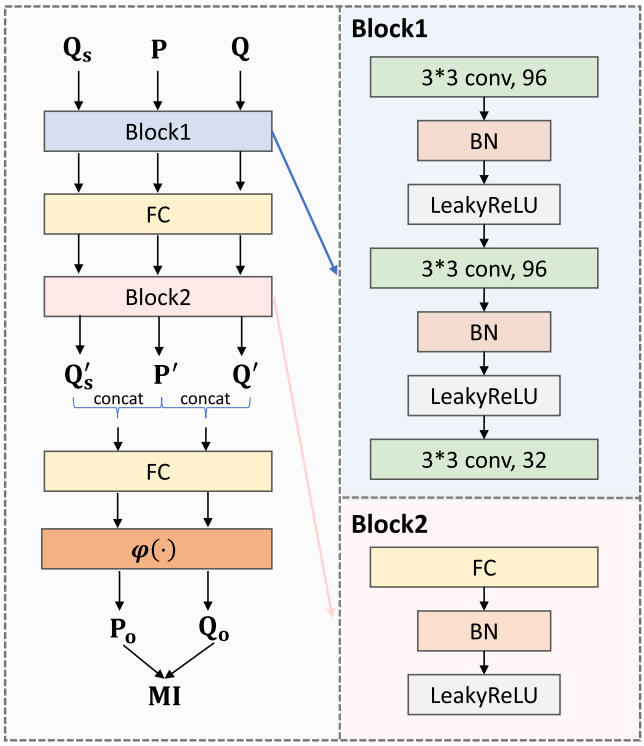


Fig. 4. Architecture of the module to calculate mutual information. FC and BN denote the fully connected layer and batch normalization, respectively. Given input feature maps $\mathbf{P}$ and $\mathbf{Q}$, we denote randomly shuffled $\mathbf{Q}$ as $\mathbf{Q}_s$. After inputting this result into Block1 and Block2, with architectures shown on the right, we obtain the intermediate result $\mathbf{P}'$, $\mathbf{Q}'$, and $\mathbf{Q}'_s$. We concatenate $\mathbf{P}'$ with $\mathbf{Q}'$ and $\mathbf{Q}'_s$ separately, and further feed the concatenated result into an FC layer. $\varphi(\cdot)$ represents the Softplus operation in Eqn. 10. The difference between $\mathbf{P}_o$ and $\mathbf{Q}_o$ represents the mutual information (MI).

whenever $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}^2$. $\mathbb{P}_{\mathbf{X}^m\hat{\mathbf{Y}}_1^m}$ and $\mathbb{P}_{\mathbf{X}^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_1^m}$ represent the joint distributions and the product of the marginals, respectively. Thus, the larger the KL-divergence between the joint distributions and the product of the marginals, the stronger the dependence between $\mathbf{X}^m$ and $\hat{\mathbf{Y}}_1^m$.

However, as Eqn. 6 is not optimizable by neural networks, we further exploit the following bound by introducing a deep neural network with parameters $\theta' \in \Theta$ [65]:

$$I(\mathbf{X}^m; \hat{\mathbf{Y}}_1^m) \geq I_\Theta(\mathbf{X}^m, \hat{\mathbf{Y}}_1^m). \tag{8}$$

Specifically, we introduce Jensen-Shannon representation (JS) as the *neural information measure* as it is stable in the optimization of neural networks. The estimated mutual information with JS is defined as:

$$I_\Theta^{JS}(\mathbf{X}^m, \hat{\mathbf{Y}}_1^m) = \sup_{\theta' \in \Theta} \mathbb{E}_{\mathbb{P}_{\mathbf{X}^m\hat{\mathbf{Y}}_1^m}}[-\varphi(-T_{\theta'})] - \mathbb{E}_{\mathbb{P}_{\mathbf{X}^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_1^m}}[\varphi(T_{\theta'})], \tag{9}$$

where $\{T_{\theta'}\}_{\theta' \in \Theta}$ denotes a set of functions parameterized by a neural network to maximize the mutual information. The supremum is taken over all functions $T_{\theta'}$ such that the two expectations are finite. The $\varphi(\cdot)$ represents the Softplus operation:

$$\varphi(x) = \log\left(1 + e^x\right). \tag{10}$$

Similarly, the mutual information between the intermediate representation $\hat{\mathbf{Y}}_1^m$ and final representation $\hat{\mathbf{Y}}_4^m$ is calculated by:

$$I_\Theta^{JS}(\hat{\mathbf{Y}}_1^m, \hat{\mathbf{Y}}_4^m) = \sup_{\theta' \in \Theta} \mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}_1^m\hat{\mathbf{Y}}_4^m}}[-\varphi(-T_{\theta'})] - \mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}_1^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_4^m}}[\varphi(T_{\theta'})]. \tag{11}$$

Thus, Eqn. 5 is then reformulated as:

$$\mathcal{L}_{IB} = \min_{\hat{\mathbf{Y}}_1^m} \sup_{\theta' \in \Theta} \mathbb{E}_{\mathbb{P}_{\mathbf{X}^m\hat{\mathbf{Y}}_1^m}}[-\varphi(-T_{\theta'})] - \mathbb{E}_{\mathbb{P}_{\mathbf{X}^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_1^m}}[\varphi(T_{\theta'})] - \beta\mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}_1^m\hat{\mathbf{Y}}_4^m}}[-\varphi(-T_{\theta'})] + \beta\mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}_1^m} \otimes \mathbb{P}_{\hat{\mathbf{Y}}_4^m}}[\varphi(T_{\theta'})], \tag{12}$$

where β is set to 1 in our case. Note that we optimize the parameters of the information bottleneck and the original Light CNN network, i.e., θ and θ', at the same time during the training phase. For concise representation, we denote the union of them as parameters θ''.

2) Search Algorithm: Inspired by neural architecture search (NAS) [25], we propose an automatic feature search (AFS) algorithm. Different from common NAS methods that deal with discrete operations, our saliency indicator is implemented in a continuous space; thus, we directly apply gradient descent in the optimization process. Let $\mathcal{L}_{train}$ and $\mathcal{L}_{val}$ denote the training and validation loss, respectively. These parameters are determined together by the saliency indicator $\mathbf{M}_B^m$ and the network weights θ''. This solution process involves a bilevel optimization problem, where $\mathbf{M}_B^m$ is the upper-level variable and θ'' is the lower-level variable:

$$\begin{aligned} &\min_{\mathbf{M}_B^m} \ \mathcal{L}_{val}\left(\theta^*\left(\mathbf{M}_B^m\right), \mathbf{M}_B^m\right) \\ &\text{s.t. } \theta^*\left(\mathbf{M}_B^m\right) = \arg\min_{\theta''} \mathcal{L}_{train}\left(\theta'', \mathbf{M}_B^m\right). \end{aligned} \tag{13}$$

As shown in Alg. 1, the network is trained in a two-stage way. The training set $\mathbb{S}_T$ is randomly split into $\mathbb{S}_{train}$ and $\mathbb{S}_{val}$, which are responsible for training and validation during the training phase, respectively. In the first stage, we optimize the saliency indicator $\mathbf{M}_B^m$ by descending $\nabla_{\mathbf{M}_B^m}\mathcal{L}_{\text{val}}\left(\theta'', \mathbf{M}_B^m\right)$. Then at the second stage, the network weight θ'' is optimized by descending $\nabla_{\theta''}\mathcal{L}_{\text{train}}\left(\theta'', \mathbf{M}_B^m\right)$. θ'' is further optimized with the searched $\mathbf{M}_B^m$ based on the trained $\mathbb{S}_T$.

C. Overall Loss Function

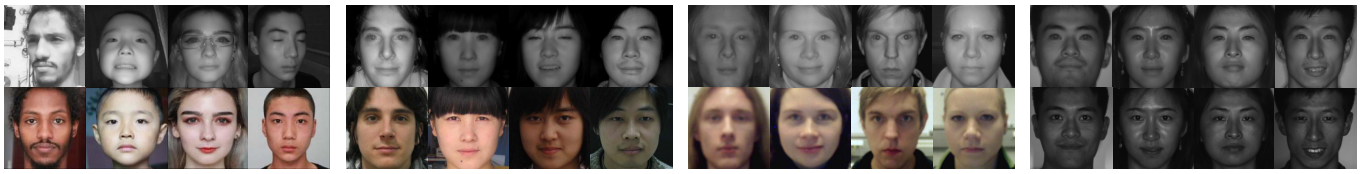
In general, the overall loss function of our proposed SSN is denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \lambda_{ib}\mathcal{L}_{ib}, \tag{14}$$

where λ_{ib} is set to 0.001.

IV. Experiments

To evaluate the performance of the proposed Saliency Search Network (SSN), we conduct extensive experiments on multiple datasets, including the CASIA NIR-VIS 2.0 Face dataset [22], the Oulu-CASIA NIR-VIS dataset [23], the BUAA-VisNir Face dataset [21], and the LAMP-HQ dataset [17]. We conduct experiments on both complete NIR face images and incomplete ones. Moreover, to make the



(a) The LAMP-HQ dataset (b) The CASIA NIR-VIS 2.0 dataset (d) The Oulu-CASIA NIR-VIS dataset (c) The BUAA-VisNir dataset

Fig. 5. Illustration of used NIR images of the four introduced datasets. The first row presents NIR probe images and the second row shows the corresponding VIS gallery images.

TABLE I

FACE RECOGNITION AND VERIFICATION RATES (%) ON THE LAMP-HQ DATASET. NOTE COMPLETE NIR FACE IMAGES ARE USED IN THESE EXPERIMENTS. THE TOP GROUP INCLUDES TRADITIONAL METHODS, AND THE BOTTOM GROUP GIVES RESULTS OF DEEP LEARNING-BASED METHODS

Method	Rank-1	VR@FPR=1%	VR@FPR=0.1%
ADFL [66]	95.1±0.5	92.1±0.9	73.3±2.2
PCFH [18]	95.3±0.5	92.9±0.6	75.1±1.8
PACH [19]	95.4±0.5	93.1±0.4	75.3±1.7
VSANet [17]	97.3±0.2	94.8±0.7	78.2±3.0
Light CNN [20]	94.6±0.3	92.5±0.6	75.6±1.9
SSN (Ours)	**98.6±0.2**	**99.1±0.2**	**97.5±0.3**

TABLE II

FACE VERIFICATION AND RECOGNITION RATES (%) ON THE CASIA NIR-VIS 2.0 DATASET. NOTE COMPLETE NIR FACE IMAGES ARE USED IN THESE EXPERIMENTS

Method	Rank-1	VR@FPR=1%	VR@FPR=0.1%
DSIFT [67]	73.3±1.1	-	-
CDFL [68]	71.5±1.4	67.7	55.1
Gabor+RBM [69]	86.2±1.0	-	81.3±1.8
H2(LBP3) [37]	43.8	36.5	10.1
CEFD [70]	85.6	-	-
Recon.+UDP [71]	78.5±1.7	85.8	-
HFR-CNN [15]	85.9±0.9	-	78
TRIVET [14]	95.7±0.5	98.1±0.3	91.0±1.3
IDNet [38]	87.1±0.9	-	74.5
ADFL [66]	98.2±0.3	99.1±0.2	97.2±0.5
Hallucination [72]	89.6±0.9	-	-
DLFace [71]	98.68	-	-
W-CNN [34]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [19]	98.9±0.2	98.3±0.2	-
RCN [73]	99.3±0.2	98.7±0.2	-
MC-CNN [74]	99.4±0.1	99.3±0.1	-
DVR [16]	99.7±0.1	99.6±0.3	98.6±0.3
LLRe-rank [75]	98.7	99.4	96.5
CFC-Fuse [76]	99.5±0.1	99.8±0.1	97.5±0.2
PCFH [18]	98.8±0.3	99.6±0.1	97.7±0.3
VSANet [17]	99.2±**0.0**	99.7±0.0	98.2±0.2
ADCANs [77]	99.1±0.2	99.6±0.1	98.5±0.2
Light CNN [20]	96.7±0.2	98.5±0.6	94.8±0.4
SSN (Ours)	**99.9**±0.1	**100.0±0.0**	**99.8±0.0**

experiments more persuasive, we use two types of occlusions, namely "maskTop" and "maskBottom" to mask the complete NIR face images. These image types are corresponding to the eyes part and lower half of the face, respectively. For each type of NIR face image, we compare our results with those of other state-of-the-art methods and provide in-depth discussion and analyses. Moreover, abundant ablation studies are performed to evaluate the effectiveness of every single component of our SSN. The details will are presented in the following subsections.

A. Datasets

The CASIA NIR-VIS 2.0 Face dataset [22] is one of the most popular datasets in the field of NIR-VIS face recognition. It involves 725 subjects with 17,580 images in total. For each subject, the number of NIR/VIS face images ranks from 1/5 to 22/50. Note that the NIR and VIS images of a specific person are randomly and asynchronously captured; thus, they are not in pairs. The images in the CASIA NIR-VIS 2.0 Face dataset differ from each other in illumination, pose, age, and so on, making it a challenging task to perform NIR-VIS face recognition on this dataset.

The authors of this dataset developed two types of protocols: View 1 for developing algorithms and View 2 for training and testing. We follow View 2 to set our experiments. Specifically, we perform a 10-fold experiment, where each fold contains distinct training and testing face images. Note that to ensure the fairness of the experiments, there is no overlap between the training and testing sets. For each fold, there are approximately 2,500 VIS face images and 6,100 NIR ones corresponding to about 360 subjects. During testing, 6,210 NIR and 358 VIS images of the other 358 subjects form the probe and gallery sets, respectively.

The Oulu-CASIA NIR-VIS dataset [23] was proposed to deal with the problems caused by illumination in face recognition. There are a total of 7,680 images corresponding to 80 subjects, where 50 of them are from Oulu University, while the rest are from CASIA. The NIR and VIS face images are captured under three illumination conditions, including normal indoor, weak, and dark lighting conditions. Six kinds of expressions are also taken into consideration, i.e., anger, disgust, fear, happiness, sadness, and surprise. Following [37], we select 40 subjects from the Oulu-CASIA NIR-VIS dataset; images for 20 individuals are used for training and the rest are used for testing. For each subject, there are 48 NIR face images and 48 VIS face images. Specifically, eight

TABLE III

FACE VERIFICATION AND RECOGNITION RATES (%) ON THE BUAA-VISNIR FACE DATASET (LEFT) AND THE OULU-CASIA NIR-VIS DATASET (RIGHT). NOTE COMPLETE NIR FACE IMAGES ARE USED IN THESE EXPERIMENTS

Method	BUAA-VisNir			Oulu-CASIA NIR-VIS		
	Rank-1	VR@FPR=1%	VR@FPR=0.1%	Rank-1	VR@FPR=1%	VR@FPR=0.1%
MPL3 [78]	53.2	58.1	33.3	48.9	41.9	11.4
KCSR [23]	81.4	83.8	66.7	66.0	49.7	26.1
KPS [79]	66.6	60.2	41.7	62.2	48.3	22.2
KDSR [80]	83	86.8	69.5	66.9	56.1	31.9
H2(LBP3) [37]	88.8	88.8	73.4	70.8	62.0	33.6
TRIVET [14]	93.9	93	80.9	92.2	67.9	33.6
ADFL [66]	95.2	95.3	88	95.5	83.0	60.7
CFC [76]	99.7	98.7	97.8	99.9	98.1	90.7
W-CNN [34]	97.4	96.0	91.9	98.0	81.5	54.6
PACH [19]	98.6	98.0	93.5	100.0	97.9	88.2
DVR [16]	99.2	98.5	96.9	100.0	97.2	84.9
VSANet [17]	98.8	98.3	93.4	100.0	97.7	89.0
PCFH [18]	**100.0**	97.7	86.6	100.0	97.7	86.6
ADCANs [77]	99.8	99.7	98.4	99.8	93.2	78.9
Light CNN [20]	96.5	95.4	86.7	96.7	92.4	65.1
SSN (Ours)	99.9	**99.9**	**99.5**	**100.0**	**99.7**	**98.1**

images are randomly selected from each expression for each domain.

The BUAA-VisNir face dataset [21] is a commonly used dataset for HFR. It contains 2,700 images from 150 subjects. They are split into two groups: 900 images of 50 subjects for training and 1800 images of the remaining 100 subjects for testing. For each subject, there are nine pairs of NIR and VIS images corresponding to nine different poses or expressions, including neutral-frontal, left-rotation, right-rotation, tilt-up, tilt-down, happiness, anger, sorrow, and surprise. Note for each pair, the NIR and VIS face images are captured simultaneously by a single multi-spectral camera.

The LAMP-HQ dataset [17] is a newly introduced large-scale dataset to address the problems involving NIR-VIS HFR. There are 56,788 NIR and 16,828 VIS face images of 573 subjects, making it much larger than the previously discussed datasets. The images are asynchronously captured, resulting in unpaired NIR and VIS face images. Moreover, the images are distinct from each other in illumination, pose, scene, glasses, expressions, and so on, further increasing the level of complexity. Following [17], both the training and testing sets account for approximately 50% of the dataset. Similar to the approach in [22], a 10-fold experimental protocol is used, where each fold uses distinct training and testing sets chosen randomly. Note that there is no overlap between the training and testing datasets.

B. Reproducibility

To begin the experiments, images from all datasets are preprocessed in the same way. First, we detect 68 keypoints of these face images and align them using MTCNN [85]. Then, all images are cropped and resized to a 128×128 resolution. The images are then used for complete NIR-VIS HFR. As shown in the left part of Fig. 6, for partial NIR-VIS HFR, images are further occluded by masks. We adopt two kinds of masks here, which are "maskTop" and "maskBottom" that corresponding to the eyes part and the lower half of the face, respectively. Specifically, if we denote the coordinate of point n as (x_n, y_n), we have "maskTop" as a rectangular with width equals to $x_{14} - x_2$, and height equals to $y_{10} - y_{30}$. The starting point is $(x_{30} - 0.5width, y_{30})$. Additionally, "maskBottom" is a bigger rectangular with a width equals to $x_{27} - x_{18}$ and height equals to $y_{10} - y_{30}$ starting from point (x_{18}, y_{18}).

SSN is developed based on the architecture of Light CNN. As shown in Fig. 3, the gray boxes denote the original Light CNN network. The dashed arrows indicate the original data flows, which are replaced by flows through PSBs. In addition, the search process is guided by an information bottleneck (IB) network. As discussed in Eqn. 5, an IB is calculated by mutual information (MI), and the detailed architecture is shown in Fig. 4.

During training, the models for complete and partial NIR-VIS HFR are trained separately. We follow View 2 in [22] to perform a 10-fold training. In partial NIR-VIS HFR, we apply masks with randomly selected sizes and positions to the input NIR face images during every iteration of the training phase. The model is further used during the testing phase for different kinds of occlusions, including "maskTop" and "maskBottom", to assess the robustness and generalization of the model. Moreover, to perform a fair comparison, all the introduced comparison methods in partial NIR-VIS HFR, including PCFH [18], PACH [19], VSANet [17], DFNet [84], and LBAM [83], are retrained with the same masking strategy. Note that we use models trained on the CASIA NIR-VIS 2.0 Face dataset to perform testing on the CASIA NIR-VIS 2.0 Face dataset, the BUAA-NirVis face dataset, and the Oulu-CASIA NIR-VIS dataset. The training and testing on the LAMP-HQ dataset are conducted individually following the settings given in [17].

We implement our networks using PyTorch. The network is trained in a two-stage manner with an NVIDIA Tesla V100S, occupying 11G GPU. During the first stage, the saliency

TABLE IV

FACE VERIFICATION AND RECOGNITION RATES (%) ON THE CASIA NIR-VIS 2.0 FACE DATASET WITH "MASKBOTTOM" AND "MASKTOP" AS THE INTRODUCED OCCLUSIONS. FOR RESULTS OF EACH KIND OF OCCLUSION, THE GROUPS FROM TOP TO BOTTOM INDICATE THE EXPERIMENTS OF SALIENCY DETECTION METHODS, NIR-VIS HFR METHODS, VISUAL FACE COMPLETION METHODS, AND FUSION-BASED METHODS. NOTE THAT THE NAME OF FUSION-BASED METHODS ARE IN TERM OF "A + B", WITH "A" AS THE NIR-VIS HFR METHOD AND "B" AS THE OCCLUDED FR METHOD. THE METHOD NAMES ENDING IN "*_ori*" AND "*_retrain*" INDICATE THE USED MODELS ARE OFFICIALLY RELEASED AND RETRAINED BY OUR MASKING STRATEGY, RESPECTIVELY

Method	AUC	EER	TPR@FPR=1%	TPR@FPR=0.1%	Rank-1	Rank-2	Rank-3	Rank-4
maskBottom								
GCPANet [81]	96.8±0.2	9.4±0.2	66.6±1.4	38.2±1.9	55.1±2.0	65.4±1.7	70.6±1.4	74.1±1.3
EGNet [82]	97.0±0.2	9.2±0.2	68.5±1.3	40.9±2.2	54.7±1.9	64.6±1.9	69.9±1.6	73.6±1.5
VSANet_ori [17]	98.1±0.1	7.0±0.3	75.5±1.2	49.3±2.5	60.9±2.2	70.9±1.5	75.8±1.2	78.9±0.9
VSANet_retrain	98.5±0.1	6.4±0.3	78.5±1.0	54.7±2.5	65.2±1.8	74.1±1.1	78.4±0.9	81.3±0.9
PCFH_ori [18]	95.7±0.2	11.2±0.4	55.3±1.9	25.1±2.3	35.7±1.7	46.1±1.6	52.3±1.8	56.7±1.9
PCFH_retrain	96.8±0.2	9.4±0.3	67.7±1.0	40.8±1.9	52.3±2.1	62.6±1.7	68.1±1.4	72.0±1.1
PACH_ori [19]	97.8±0.1	7.5±0.3	72.6±1.2	45.5±2.5	57.3±2.0	68.1±1.4	73.0±1.1	76.2±0.8
PACH_retrain	98.2±0.3	7.0±0.7	72.5±4.6	45.6±7.5	57.0±5.5	65.7±4.7	70.6±4.2	74.0±3.7
LBAM_retrain [83]	97.5±0.2	8.3±0.4	73.3±1.4	49.1±2.3	58.8±1.6	68.4±1.6	73.6±1.6	77.0±1.6
DFNet_retrain [84]	99.1±0.1	4.8±0.3	86.1±0.7	61.6±2.4	69.9±1.7	78.7±1.4	82.8±1.1	85.6±1.0
VSANet [17] + LBAM [83]	99.8±0.1	5.3±0.2	81.9±1.0	55.6±2.0	66.8±0.7	76.3±0.7	80.7±0.8	83.5±0.9
PCFH [18] + LBAM [83]	97.2±0.2	8.6±0.3	67.8±1.9	38.4±2.1	49.4±2.0	60.7±2.0	66.4±1.6	69.9±1.3
PACH [19] + LBAM [83]	98.0±0.1	7.1±0.2	74.2±0.7	44.7±2.7	57.3±1.8	68.3±1.2	73.5±0.9	76.9±0.8
VSANet [17] + DFNet [84]	98.9±0.1	5.0±0.3	83.9±1.0	58.9±1.8	68.5±1.1	77.6±1.1	81.9±1.1	84.8±1.1
PCFH [18] + DFNet [84]	95.8±0.3	10.9±0.5	60.7±1.2	33.3±2.1	44.1±1.6	55.0±1.4	61.0±1.5	65.2±1.3
PACH [19] + DFNet [84]	96.9±0.2	9.0±0.3	67.1±1.0	38.9±3.0	49.5±2.3	61.2±1.5	66.8±1.2	70.5±1.0
Light CNN [20]	97.7±0.1	7.9±0.3	72.3±1.1	44.7±2.3	57.7±2.2	68.2±1.4	73.1±1.1	76.2±1.0
Ours	**99.8±0.0**	**2.1±0.3**	**96.6±0.6**	**86.5±0.7**	**90.8±1.0**	**94.9±0.7**	**96.3±0.7**	**97.1±0.5**
maskTop								
GCPANet [81]	99.6±0.0	3.3±0.2	91.4±0.7	73.3±1.3	82.0±1.3	88.9±1.2	91.8±0.9	93.6±0.7
EGNet [82]	99.5±0.0	3.5±0.2	90.1±0.7	69.9±1.6	79.9±1.3	87.3±1.2	90.7±1.0	92.7±0.8
VSANet_ori [17]	99.8±0.0	1.9±0.2	96.4±0.4	86.8±1.4	91.1±0.8	95.1±0.6	96.5±0.5	97.3±0.5
VSANet_retrain	99.8±0.0	1.7±0.2	97.1±0.5	88.9±1.3	92.7±1.2	96.1±0.7	97.3±0.5	97.9±0.4
PCFH_ori [18]	99.0±0.1	5.1±0.3	82.7±1.0	56.0±2.3	67.5±1.4	76.6±1.3	80.9±1.1	83.8±1.0
PCFH_retrain	99.4±0.1	3.8±0.3	89.2±0.8	69.2±1.7	77.7±1.5	85.5±1.4	88.9±1.0	90.9±0.8
PACH_ori [19]	99.4±0.1	3.8±0.3	89.9±0.6	71.6±1.1	81.4±0.9	88.4±1.0	91.3±0.9	92.9±0.8
PACH_retrain	99.1±0.2	4.8±0.5	85.1±2.3	62.8±5.1	73.4±3.2	80.7±2.9	84.1±2.4	86.3±2.1
LBAM_retrain [83]	99.1±0.1	4.9±0.4	83.6±1.2	54.2±2.9	67.2±1.7	77.4±1.5	82.1±1.4	85.2±1.2
DFNet_retrain [84]	99.7±0.0	2.5±0.2	95.4±0.5	84.9±1.2	89.2±1.0	93.3±0.7	94.8±0.6	95.7±0.6
VSANet [17] + LBAM [83]	99.8±0.0	1.9±0.2	96.9±0.5	88.0±0.8	91.7±1.0	95.4±0.7	96.8±0.6	97.5±0.5
PCFH [18] + LBAM [83]	99.5±0.1	3.3±0.2	90.7±0.8	69.4±1.8	80.8±1.5	87.8±1.1	90.9±1.2	92.8±1.0
PACH [19] + LBAM [83]	99.6±0.0	2.9±0.1	93.5±0.5	77.9±0.8	85.8±0.8	91.6±0.7	94.0±0.6	95.3±0.5
VSANet [17] + DFNet [84]	99.9±0.0	1.4±0.2	97.1±0.4	88.9±1.2	93.8±1.1	96.7±0.8	97.7±0.5	98.3±0.4
PCFH [18] + DFNet [84]	99.7±0.0	2.7±0.2	93.6±0.7	76.2±1.6	83.7±1.0	90.0±0.9	92.7±0.8	94.0±0.7
PACH [19] + DFNet [84]	99.8±0.0	2.2±0.1	95.6±0.5	82.8±0.9	88.1±0.9	93.5±0.6	95.4±0.5	96.5±0.5
Light CNN [20]	99.7±0.0	2.6±0.2	93.9±0.7	78.8±1.2	85.6±1.1	91.4±1.0	93.7±0.8	95.1±0.7
Ours	**100.0±0.0**	**0.6±0.0**	**99.7±0.1**	**97.2±0.7**	**98.4±0.3**	**99.4±0.2**	**99.7±0.1**	**99.8±0.1**

indicator $\mathbf{M}_B^m$ is updated by the AFS algorithm with an Adam optimizer at a learning rate of 0.001. Then the network weights θ'' is updated in the second stage. We introduce an SGD optimizer with a learning rate of 0.0001. Our model is trained for 12 epochs and 40 epochs for the first and second stages, respectively. Since the structure of the framework keeps consistent, the time complexity brought by the network structure is $O(1)$. Following [25], the time complexity brought by the AFS algorithm is $O(|\theta''| + |\mathbf{M}_B^m|)$, where θ'' and $\mathbf{M}_B^m$ denote the network weights and the saliency indicator, respectively. The computational demand is 3.64 GFLOPS in one forward modeling cycle. It takes about 13 minutes to train an epoch. The inference time is about 0.03 seconds per image. We set the batch size to 128.

C. Complete NIR-VIS HFR Analyses

For complete NIR-VIS HFR, we present the recognition and verification accuracies achieved by traditional methods and deep learning-based methods. Specifically, Table II presents the results of experiments on the CASIA NIR-VIS 2.0 Face dataset. We provide the results of DSIFT [67], Coupled Discriminant Feature Learning (CDFL) [68], Gabor+RBM [69], H2(LBP3) [37], Common Encoding Feature Discriminant (CEFD) [70], and Recon.+UDP [71], which are traditional methods. Most traditional methods achieve Rank-1 accuracy that is lower than 90%, which is below the required level in most practical applications. Later, with the development of deep learning, methods based on neural networks have

TABLE V

FACE VERIFICATION AND RECOGNITION RATES (%) ON THE LAMP-HQ DATASET WITH "MASKBOTTOM" AND "MASKTOP" AS THE INTRODUCED OCCLUSIONS. FACE VERIFICATION AND RECOGNITION RATES (%) ON THE CASIA NIR-VIS 2.0 FACE DATASET WITH "MASKBOTTOM" AND "MASKTOP" AS THE INTRODUCED OCCLUSIONS. FOR RESULTS OF EACH KIND OF OCCLUSION, THE GROUPS FROM TOP TO BOTTOM INDICATE THE EXPERIMENTS OF SALIENCY DETECTION METHODS, NIR-VIS HFR METHODS, VISUAL FACE COMPLETION METHODS, AND FUSION-BASED METHODS. NOTE THAT THE NAME OF FUSION-BASED METHODS ARE IN TERM OF "A + B", WITH "A" AS THE NIR-VIS HFR METHOD AND "B" AS THE OCCLUDED FR METHOD. THE METHOD NAMES ENDING IN "*_ori*" AND "*_retrain*" INDICATE THE USED MODELS ARE OFFICIALLY RELEASED AND RETRAINED BY OUR MASKING STRATEGY, RESPECTIVELY

Method	AUC	EER	TPR@FPR=1%	TPR@FPR=0.1%	Rank-1	Rank-2	Rank-3	Rank-4
maskBottom								
GCPANet [81]	94.8±0.3	12.5±0.5	56.3±1.5	29.3±1.6	53.3±1.4	63.8±1.4	69.3±1.4	73.0±1.4
EGNet [82]	94.1±0.3	13.5±0.4	55.2±1.4	30.3±1.4	51.3±1.6	61.8±1.4	67.3±1.4	71.0±1.4
VSANet_ori [17]	95.9±0.2	11.1±0.4	62.2±1.3	34.7±1.7	58.1±1.5	68.3±1.2	73.5±1.2	76.8±1.2
VSANet_retrain	96.6±0.2	9.8±0.3	62.5±1.4	33.9±2.0	63.5±1.3	73.2±1.2	77.9±1.2	81.0±1.1
PCFH_ori [18]	94.5±0.3	13.0±0.4	56.2±1.5	30.6±1.5	51.7±1.3	62.3±1.3	67.9±1.4	71.6±1.3
PCFH_retrain	95.1±0.3	12.2±0.4	57.8±1.3	30.0±1.8	52.5±1.4	63.2±1.3	68.9±1.2	72.7±1.1
PACH_ori [19]	95.3±0.3	11.9±0.4	59.5±1.3	32.2±1.7	54.0±1.6	64.3±1.5	69.7±1.5	73.3±1.4
PACH_retrain	97.3±0.2	8.6±0.3	66.8±1.0	37.2±1.7	68.5±1.3	77.9±1.2	82.3±1.1	85.0±1.0
LBAM_retrain [83]	94.8±0.3	12.5±0.4	50.3±1.4	22.5±1.0	49.1±1.1	60.0±1.1	66.0±1.1	70.1±1.1
DFNet_retrain [84]	95.8±0.2	10.9±0.3	55.0±1.2	27.1±1.5	57.5±1.0	67.7±1.1	73.0±1.0	76.5±1.0
VSANet [17] + LBAM [83]	96.6±0.2	9.7±0.3	64.9±0.9	36.2±1.4	62.5±1.6	72.4±1.5	77.2±1.4	80.2±1.3
PCFH [18] + LBAM [83]	94.2±0.3	13.1±0.4	53.5±1.3	27.0±1.1	50.6±1.6	60.8±1.6	66.4±1.5	70.2±1.5
PACH [19] + LBAM [83]	95.6±0.2	11.3±0.3	59.2±1.2	31.3±1.3	56.6±1.7	66.6±1.6	71.9±1.4	75.3±1.4
VSANet [17] + DFNet [84]	96.9±0.2	9.4±0.3	66.3±0.8	39.0±1.2	65.1±1.0	74.7±1.0	79.3±1.0	82.2±1.0
PCFH [18] + DFNet [84]	94.0±0.3	13.7±0.4	52.2±1.3	26.4±1.3	47.1±1.7	57.6±1.6	63.3±1.5	67.1±1.4
PACH [19] + DFNet [84]	95.7±0.2	11.3±0.4	60.3±1.2	32.2±1.5	56.6±1.4	67.2±1.2	72.7±1.1	76.2±1.1
Light CNN [20]	95.7±0.3	11.2±0.4	61.3±1.4	34.3±1.7	58.0±1.3	68.2±1.3	73.4±1.3	76.8±1.3
Ours	**99.7±0.0**	**2.3±0.2**	**95.6±0.6**	**83.8±1.2**	**91.7±0.9**	**95.5±0.6**	**96.8±0.5**	**97.5±0.3**
maskTop								
GCPANet [81]	96.8±0.3	9.5±0.5	68.3±0.9	43.0±1.5	57.7±1.3	67.0±1.1	71.5±1.1	74.5±1.0
EGNet [82]	96.1±0.3	10.3±0.4	63.9±1.2	38.3±1.9	54.3±1.9	63.8±1.5	68.6±1.3	71.8±1.2
VSANet_ori [17]	97.4±0.3	8.3±0.4	71.2±0.6	46.1±1.1	62.7±1.2	72.0±1.0	76.4±0.9	79.1±0.8
VSANet_retrain	98.2±0.2	6.7±0.4	75.9±1.0	49.7±1.8	67.8±1.4	76.7±1.2	80.9±1.2	83.4±1.1
PCFH_ori [18]	96.7±0.3	9.7±0.5	66.2±1.0	40.2±1.4	55.8±1.4	65.4±1.3	70.2±1.3	73.4±1.2
PCFH_retrain	97.3±0.3	8.6±0.4	69.0±0.7	41.9±1.5	59.5±1.0	69.2±0.9	74.0±0.9	77.1±0.8
PACH_ori [19]	97.1±0.3	9.0±0.4	68.0±0.5	41.1±1.6	57.0±1.2	66.8±1.0	71.5±0.9	74.7±0.8
PACH_retrain	98.5±0.2	5.5±0.3	78.9±1.4	49.2±2.0	74.1±1.0	82.5±1.0	86.3±1.0	88.7±0.9
LBAM_retrain [83]	96.3±0.2	10.2±0.4	58.5±0.8	29.9±1.2	51.8±0.7	61.8±0.7	67.1±0.8	70.7±0.8
DFNet_retrain [84]	96.5±0.2	9.9±0.3	58.6±1.1	31.4±1.1	55.9±0.8	65.7±0.8	70.8±0.8	74.3±0.8
VSANet [17] + LBAM [83]	98.3±0.2	6.5±0.3	77.9±0.9	52.4±0.9	71.4±0.7	79.9±0.7	83.6±0.7	85.9±0.7
PCFH [18] + LBAM [83]	96.7±0.3	9.5±0.5	63.1±1.0	34.0±2.1	52.2±1.8	62.3±1.4	67.5±1.2	71.0±1.1
PACH [19] + LBAM [83]	97.5±0.2	8.2±0.4	68.7±0.8	40.1±1.9	57.4±1.4	67.6±1.3	72.7±1.1	76.0±1.0
VSANet [17] + DFNet [84]	98.4±0.2	6.3±0.3	79.9±0.6	55.4±1.0	71.2±0.6	79.6±0.6	83.3±0.6	85.6±0.6
PCFH [18] + DFNet [84]	96.4±0.3	10.1±0.5	61.5±0.7	32.5±1.5	50.2±0.8	60.5±0.7	66.0±0.7	69.7±0.7
PACH [19] + DFNet [84]	97.6±0.2	8.0±0.4	71.5±0.5	44.0±1.6	60.7±1.0	70.5±0.8	75.2±0.8	78.3±0.7
Light CNN [20]	97.6±0.3	8.0±0.5	72.5±0.8	47.5±1.4	61.7±1.2	71.0±1.0	75.5±1.0	78.3±0.9
Ours	**99.7±0.1**	**1.7±0.2**	**97.4±0.5**	**89.1±1.1**	**94.5±0.6**	**97.2±0.4**	**98.1±0.4**	**98.5±0.3**

become mainstream. We provide results of recently proposed HFR-CNN [15], TRIVET [14], IDNet [38], Adversarial Discriminative Feature Learning (ADFL) [66], Hallucination [72], DLFace [71], W-CNN [34], Pose agnostic crossspectral hallucination (PACH) [19], Residual compensation networks (RCN) [73], MC-CNN [74], DVR [16], LLRe-rank [75], CFC-Fuse [76], PCFH [18], VSANet [17], and ADCANs [77]. The result of Light CNN [20] is provided as the baseline of our network. Note that the Rank-1 accuracy and verification results of our method are significantly better than those of the Light CNN, indicating that the proposed SSN outperforms the baseline model. Additionally, our SSN outperforms most of the existing methods in both accuracy and stability. Although the standard deviation ahieved by VSANet is better than that of our model, considering the accuracy improvement, the difference of 0.1 is acceptable.

Table I shows the results on the LAMP-HQ dataset. The proposed method yields much higher Rank-1 and verification accuracy values than the recently proposed state-of-the-art methods. Specifically, as the False Positive Rate (FPR) decreases, the verification results sufficiently indicate the superiority of our methods. As shown in Table III, the results on the BUAA-VisNir face dataset and the Oulu-CASIA NIR-VIS dataset are also provided. Apart from the methods discussed in the previous paragraph, we also introduce the MPL3 [78], KCSR [23], KPS [79], and KDSR [80] for comparison.

TABLE VI

FACE VERIFICATION AND RECOGNITION RATES (%) ON THE BUAA-VISNIR FACE DATASET AND OULU-CASIA NIR-VIS DATASET WITH "MASKBOTTOM" AND "MASKTOP" AS THE INTRODUCED OCCLUSIONS. FACE VERIFICATION AND RECOGNITION RATES (%) ON THE CASIA NIR-VIS 2.0 FACE DATASET WITH "MASKBOTTOM" AND "MASKTOP" AS THE INTRODUCED OCCLUSIONS. FOR RESULTS OF EACH KIND OF OCCLUSION, THE GROUPS FROM TOP TO BOTTOM INDICATE THE EXPERIMENTS OF SALIENCY DETECTION METHODS, NIR-VIS HFR METHODS, VISUAL FACE COMPLETION METHODS, AND FUSION-BASED METHODS. NOTE THAT THE NAME OF FUSION-BASED METHODS ARE IN TERM OF "A + B", WITH "A" AS THE NIR-VIS HFR METHOD AND "B" AS THE OCCLUDED FR METHOD. THE METHOD NAMES ENDING IN "*_ori*" AND "*_retrain*" INDICATE THE USED MODELS ARE OFFICIALLY RELEASED AND RETRAINED BY OUR MASKING STRATEGY, RESPECTIVELY

Method	BUAA-VisNir				Oulu-CASIA NIR-VIS			
	AUC	TPR@FPR=1%	TPR@FPR=0.1%	Rank-1	AUC	TPR@FPR=1%	TPR@FPR=0.1%	Rank-1
maskBottom								
GCPANet [81]	97.3	79.8	58.0	78.6	95.7	51.3	22.2	83.1
EGNet [82]	97.4	79.8	59.1	80.0	95.1	54.0	16.3	78.3
VSANet_ori [17]	97.9	83.1	61.5	81.7	96.4	58.2	30.7	88.6
VSANet_retrain	98.4	87.4	72.0	86.5	95.8	60.8	36.6	82.5
PCFH_ori [18]	97.4	78.4	54.8	78.0	89.6	39.1	18.2	59.0
PCFH_retrain	95.9	68.0	43.1	64.3	95.7	54.0	24.5	83.2
PACH_ori [19]	95.8	71.2	43.5	69.1	95.7	55.4	29.1	80.7
PACH_retrain	96.4	68.3	41.7	62.7	92.6	36.1	17.1	61.4
LBAM_retrain [83]	95.3	58.7	31.8	56.1	90.3	28.8	5.7	56.1
DFNet_retrain [84]	97.4	71.1	36.9	63.9	91.5	40.0	16.3	60.4
VSANet [17] + LBAM [83]	98.0	81.9	59.9	79.9	96.2	61.9	31.9	83.2
PCFH [18] + LBAM [83]	97.7	76.9	53.3	74.6	93.6	39.2	12.8	62.4
PACH [19] + LBAM [83]	97.7	73.8	46.5	72.0	95.5	55.3	30.3	75.3
VSANet [17] + DFNet [84]	98.4	86.7	69.1	85.4	96.0	53.5	27.4	83.8
PCFH [18] + DFNet [84]	97.6	77.9	55.8	77.1	91.6	24.0	10.6	49.4
PACH [19] + DFNet [84]	98.1	83.3	60.2	79.8	94.6	45.4	18.7	70.4
Light CNN [20]	97.9	80.3	59.1	79.4	96.4	60.0	30.4	84.6
Ours	**99.3**	**92.1**	**84.0**	**93.6**	**98.5**	**74.7**	**45.5**	**92.9**
maskTop								
GCPANet [81]	97.5	83.2	66.4	85.8	96.1	67.4	53.7	92.1
EGNet [82]	97.5	82.5	63.1	85.8	95.9	69.6	46.6	87.0
VSANet_ori [17]	97.0	79.2	55.4	80.5	96.1	71.4	41.2	91.1
VSANet_retrain	98.0	82.4	63.1	81.4	96.9	74.7	41.8	92.9
PCFH_ori [18]	97.1	76.6	55.3	76.3	93.2	56.6	34.3	77.3
PCFH_retrain	97.4	81.9	65.0	84.8	93.8	66.0	45.8	81.8
PACH_ori [19]	97.4	82.0	66.5	86.5	96.0	69.0	41.1	90.2
PACH_retrain	96.3	72.4	50.7	71.3	95.2	66.3	30.4	84.1
LBAM_retrain [83]	95.2	61.9	35.7	63.8	94.1	55.9	32.3	83.8
DFNet_retrain [84]	97.3	78.7	58.0	78.3	93.8	51.0	23.8	75.4
VSANet [17] + LBAM [83]	98.4	88.0	73.1	90.8	97.5	68.4	32.0	87.2
PCFH [18] + LBAM [83]	97.5	76.5	53.7	79.1	95.8	58.9	21.0	86.1
PACH [19] + LBAM [83]	97.7	80.6	60.9	82.8	96.2	66.6	32.2	88.6
VSANet [17] + DFNet [84]	98.9	91.7	80.6	93.8	96.9	74.3	43.3	88.2
PCFH [18] + DFNet [84]	97.4	80.3	60.4	82.2	96.0	55.7	26.0	87.0
PACH [19] + DFNet [84]	97.8	85.7	70.7	88.7	96.3	71.0	39.6	89.6
Light CNN [20]	98.0	85.8	73.4	89.1	96.1	74.0	50.6	90.3
Ours	**99.8**	**96.5**	**89.7**	**97.3**	**98.8**	**83.8**	**68.0**	**97.7**

Although some of our results are 0.1 lower than those of existing methods, our method achieves better overall performance in most of the experiments.

D. Partial NIR-VIS HFR Analyses

To perform in-depth analyses, we introduce correlation methods that belong to four categories. The first ones are saliency detection methods GCPANet [81] and EGNet [82]. We first apply their models to generate the corresponding saliency map of the input occluded NIR face images. Then we multiply the saliency maps with input images to keep only the salient area of each face. These faces are then fed into LightCNN to get the recognition and verification results. The second ones are complete NIR-VIS methods, including PCFH [18], PACH [19], and VSANet [17]. The third ones, including DFNet [84] and LBAM [83], pay attention to the partial face in the VIS domain. We conduct experiments with both their officially released models and retrained models with our masking strategy. As shown in Fig. 6, even with our training strategy, most of them still suffer from occlusions and produce face images with obvious artifacts. Among these methods, DFNet produces face images of the best quality. However, these images cannot preserve the identity

TABLE VII

ABLATION STUDY. FACE VERIFICATION AND RECOGNITION RATES (%) ON THE CASIA NIR-VIS 2.0 FACE DATASET WITH "MASKBOTTOM" AS THE INTRODUCED OCCLUSION. SI, SI (SPLIT), AND IB INDICATE THE SALIENCY INDICATOR WITHOUT CONSIDERATION OF MODALITY, THE SALIENCY INDICATOR CONSIDERING DIFFERENT MODALITIES, AND THE INFORMATION BOTTLENECK NETWORK, RESPECTIVELY. NOTE **OURS** IS EQUIPPED WITH THE AFS ALGORITHM COMPARED WITH "*w/* SI (SPILT) + IB"

Method	AUC	EER	TPR@FPR=1%	TPR@FPR=0.1%	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
Light CNN	97.7±0.1	7.9±0.3	72.3±1.1	44.7±2.3	57.7±2.2	68.2±1.4	73.1±1.1	76.2±1.0	78.5±0.9
Light CNN Finetune	98.8±0.2	5.4±0.3	79.6±3.8	47.6±7.1	65.7±3.3	74.9±2.7	79.5±2.3	82.4±2.2	84.5±1.9
w/ SI	99.6±0.1	3.0±0.5	93.3±2.1	79.9±4.4	86.6±1.6	91.8±1.0	93.9±0.9	95.0±0.9	95.6±0.8
w/ SI (split)	99.6±0.1	2.9±0.5	93.5±1.0	81.0±2.1	86.9±1.2	92.0±0.8	94.0±0.6	95.0±0.6	95.7±0.6
w/ SI (split) + IB	99.6±0.1	2.8±0.3	94.1±0.9	81.4±1.4	87.5±1.1	92.5±1.0	94.3±0.8	95.4±0.6	96.0±0.6
Ours	**99.8±0.0**	**2.1±0.3**	**96.6±0.6**	**86.5±0.7**	**90.8±1.0**	**94.9±0.7**	**96.3±0.7**	**97.1±0.5**	**97.7±0.5**

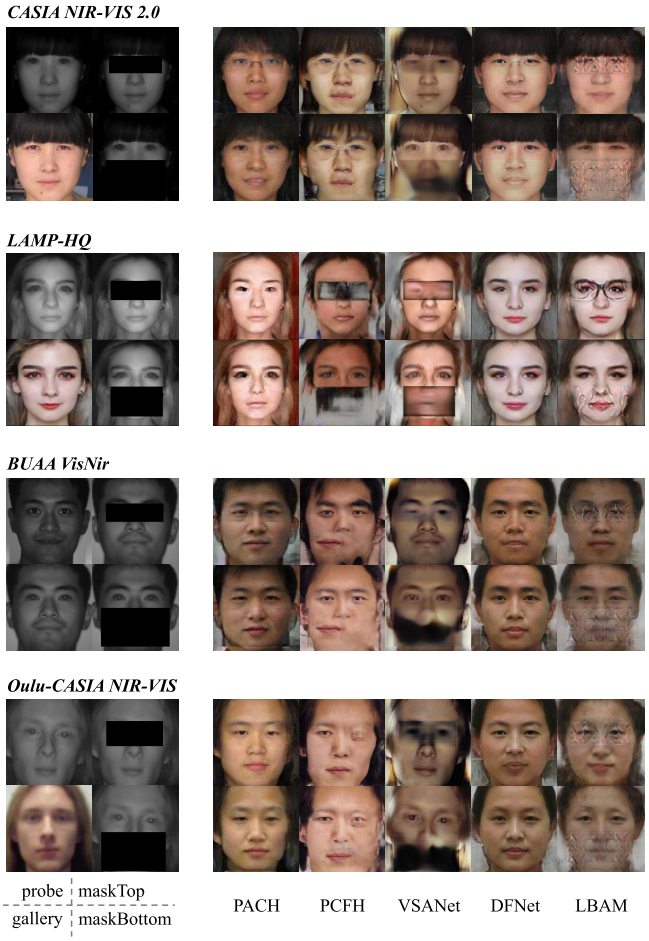


Fig. 6. Synthesis results on the CASIA NIR-VIS 2.0 Face dataset [22], the LAMP-HQ dataset [17], the BUAA-VisNir face dataset [21], and the Oulu-CASIA NIR-VIS dataset [23]. For experiments on each dataset, the left ones are input probe images with "maskTop" and "maskBottom" occlusions, as well as the corresponding gallery images. The right part of the figure presents the synthesis results of the retrained PCFH [18], PACH [19], DFNet [84], VSANet [17], and LBAM [83]. Note for each dataset, the first row and second row show the completion results from probe images with the "maskTop" and "maskBottom" occlusions, respectively.

information well, resulting in reduced face recognition performance. Moreover, we combine methods from the aforementioned two categories and perform experiments with the fusion-based methods, where we first apply NIR-VIS methods to transfer the occluded NIR face images to the VIS domain, and then utilize the impainting models to predict the missing area for further face recognition.

The quantitative results of these methods on the CASIA NIR-VIS 2.0 Face dataset, the LAMP-HQ dataset, the BUAA-VisNir face dataset, and the Oulu-CASIA NIR-VIS dataset are shown in Table IV, Table V, and Table VI, respectively. We introduce both the "maskBottom" and "maskTop" occlusions. Note that for the BUAA-VisNir face dataset and the Oulu-CASIA NIR-VIS dataset, we directly use first-fold models that trained on the CASIA NIR-VIS 2.0 Face dataset; consequently, finetuning is not needed in this case. We introduce equal error rate (EER), area under curve (AUC), and true positive rate (TPR) as the metrics for evaluating face verification performance. For most of the methods, the face verification and recognition results increase after retrained with the proposed masking strategy. However, the results of our proposed SSN surpass those of others by a large margin. For example, for Rank-1 accuracy on the CASIA NIR-VIS 2.0 Face dataset with "maskBottom" as the occlusion, we achieve an accuracy higher than 90% while other methods only achieve that lower than 70%. The significant performance improvements are attributed to the information selection ability of our model. When most other methods are trying to empower their models with the ability to predict more information, we focus on maximizing the value of the given information. Although saliency detection methods hold similar merits as ours. They are still not suitable for this task. On the one hand, the saliency maps they produced force the model to make 0/1 choices, i.e., to choose or not, which can not reflect the fine-grained importance of each pixel. On the other hand, experiments show us that it is difficult for these models to precisely recognize the occluded areas. Both the quantitative and qualitative results supply strong evidence to prove the superiority of our method.

Ablation Studies: As shown in Table VII, we also conduct ablation studies to evaluate each component of the proposed SSN, including the saliency indicator, the information bottleneck, and the AFS algorithm. The introduction of the saliency indicator improves the model performance by a large margin. For example, the Rank-1 accuracy increases to 86.6±1.6 from 65.7±3.3, which is a 31.8% increase with reference to the results of the baseline. Applying different indicators according to different modalities also contributes to the performance

improvements. In addition, the introduction of the information bottleneck and AFS algorithm further increase both the verification and recognition accuracies, indicating that each module contributes to overall model performance.

V. Conclusion

In this paper, we have proposed a modality-aware Saliency Search Network (SSN) based on Light CNN to extract domain-invariant identity features. We enabled the active parts for face recognition and disabled the inactive regions. The saliency search process is implemented by our proposed automatic feature search (AFS), where each pixel is automatically selected based on its importance for the final identity-related embedding extraction. Moreover, considering that the datasets used for NIR-VIS HFR are generally small, we further introduced an information bottleneck network to guide the search process to avoid the overfitting problem. We conducted extensive experiments involving both complete and partial NIR-VIS HFR on four well-known datasets. In-depth analyses have been made to demonstrate both the superiority of our proposed SSN over other state-of-the-art methods and the efficiency of every introduced component.

In the future, we would like to establish a new dataset of real-world occluded NIR face images, with various masks, sunglasses, etc. Our framework will then be adapted to be more suitable for real-world scenarios. Moreover, the framework and strategy can be generalized for NIR-VIS person re-identification, which is also an interesting and challenging area.

Acknowledgment

The authors would like to greatly thank the associate editor and the reviewers for their valuable comments and advice.

References

[1] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10823–10832.

[2] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, "FA-GAN: Face augmentation GAN for deformation-invariant face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2341–2355, 2021.

[3] J. Zhao, "Deep learning for human-centric image analysis," Ph.D. dissertation, Learn. Vis. Group, Fac. Eng. (FOE), Dept. Elect. Comput. Eng. (ECE), Nat. Univ. Singapore, Singapore, 2018.

[4] J. Zhao, J. Li, F. Zhao, S. Yan, and J. Feng, "Marginalized CNN: Learning deep invariant representations," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–12.

[5] J. Zhao *et al.*, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 1–3.

[6] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2207–2216.

[7] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.

[8] J. Zhao *et al.*, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 9251–9258, Jul. 2019.

[9] J. Zhao *et al.*, "Multi-prototype networks for unconstrained set-based face recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 1–13.

[10] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 460–478, Feb. 2020.

[11] J. Zhao, S. Yan, and J. Feng, "Towards age-invariant face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 23, 2020, doi: 10.1109/TPAMI.2020.3011426.

[12] J. Guo, X. Zhu, Z. Lei, and S. Z. Li, "Decomposed meta batch normalization for fast domain adaptation in face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3082–3095, 2021.

[13] W. Hu and H. Hu, "Dual adversarial disentanglement and deep representation decorrelation for NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 70–85, 2021.

[14] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for NIR-VIS heterogeneous face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[15] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. Cof. Comput. Vis. (ECCV)*, 2016, pp. 483–491.

[16] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun, "Disentangled variational representation for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 9005–9012.

[17] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He, "LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1467–1483, May 2021.

[18] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross spectral face hallucination," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 1018–1024.

[19] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7930–7938.

[20] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[21] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, 2012.

[22] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 348–353.

[23] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 156–163.

[24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 87–102.

[25] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–13.

[26] J. Zhao *et al.*, "3D-aided deep pose-invariant face recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 1–11.

[27] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Towards high fidelity face frontalization in the wild," *Int. J. Comput. Vis.*, vol. 128, pp. 1485–1504, Oct. 2019.

[28] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. NIPS*, 2018, pp. 2867–2877.

[29] C. Peng, N. Wang, J. Li, and X. Gao, "DLFace: Deep local descriptor for cross-modality face recognition," *Pattern Recognit.*, vol. 90, pp. 161–171, Jun. 2019.

[30] C. Peng, N. Wang, J. Li, and X. Gao, "Universal face photo-sketch style transfer via multiview domain translation," *IEEE Trans. Image Process.*, vol. 29, pp. 8519–8534, 2020.

[31] R. He, Y. Li, X. Wu, L. Song, Z. Chai, and X. Wei, "Coupled adversarial learning for semi-supervised heterogeneous face recognition," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107618.

[32] J. Gui and P. Li, "Multi-view feature selection for heterogeneous face recognition," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 983–988.

[33] Y. Jin, J. Li, C. Lang, and Q. Ruan, "Multi-task clustering ELM for VIS-NIR cross-modal feature learning," *Multidimensional Syst. Signal Process.*, vol. 28, no. 3, pp. 905–920, Jul. 2017.

[34] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

[35] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for cross-modal face recognition," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 426–438, May 2017.

[36] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. Int. Conf. Biometrics (ICB)*, 2009, pp. 209–218.

[37] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–463, Feb. 2017.

[38] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 54–62.

[39] H. Du, H. Shi, Y. Liu, D. Zeng, and T. Mei, "Towards NIR-VIS masked face recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 768–772, 2021.

[40] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2892–2900.

[41] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, Feb. 2018.

[42] S. Park, H. Lee, J.-H. Yoo, G. Kim, and S. Kim, "Partially occluded facial image retrieval based on a similarity measurement," *Math. Problems Eng.*, vol. 2015, Jul. 2015, Art. no. 217568.

[43] H. J. Oh, K. M. Lee, and S. U. Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image Vis. Comput.*, vol. 26, no. 11, pp. 1515–1523, 2008.

[44] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 2009, pp. 1050–1057.

[45] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren, "Structured sparse error coding for face recognition with occlusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1889–1900, May 2013.

[46] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 442–447.

[47] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential Siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 773–782.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[50] T. Elsken *et al.*, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.

[51] F. Liang *et al.*, "Computation reallocation for object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–12.

[52] W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, and Z. Wang, "FasterSEG: Searching for faster real-time semantic segmentation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–14.

[53] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 4780–4789.

[54] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*. [Online]. Available: https://arxiv.org/abs/physics/0004057

[55] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–19.

[56] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–15.

[57] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–18.

[58] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, "Learning efficient multi-agent communication: An information bottleneck approach," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9908–9918.

[59] A. Goyal *et al.*, "Infobot: Transfer and exploration via the information bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–4.

[60] M. Igl *et al.*, "Generalization in reinforcement learning with selective noise injection and information bottleneck," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1–16.

[61] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Graph information bottleneck for subgraph recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–13.

[62] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–27.

[63] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6778–6787.

[64] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1319–1327.

[65] M. I. Belghazi *et al.*, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 531–540.

[66] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7355–7362.

[67] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1788–1793.

[68] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[69] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–7.

[70] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.

[71] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 141–150.

[72] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 6628–6637.

[73] Z. Deng, X. Peng, and Y. Qiao, "Residual compensation networks for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8239–8246.

[74] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3102–3114, Jun. 2019.

[75] C. Peng, N. Wang, J. Li, and X. Gao, "Re-ranking high-dimensional deep local representation for NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4553–4565, Sep. 2019.

[76] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Cross-spectral face completion for NIR-VIS heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2019.

[77] W. Hu and H. Hu, "Adversarial disentanglement spectrum variations and cross-modality attention networks for NIR-VIS face recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 145–160, 2021.

[78] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.

[79] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1123–1128.

[80] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[81] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 10599–10606.

[82] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2019, pp. 8779–8788.

[83] C. Xie *et al.*, "Image inpainting with learnable bidirectional attention maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8858–8867.

[84] X. Hong, P. Xiong, R. Ji, and H. Fan, "Deep fusion network for image completion," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2033–2042.
[85] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

Mandi Luo (Member, IEEE) received the B.E. degree in automation engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the B.Sc. and M.Sc. degrees in electronic engineering from Katholieke Universiteit te Leuven, Leuven, Belgium, in 2017 and 2018, respectively. She is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences, Beijing, China. Her research interests include biometrics, pattern recognition, and computer vision.

Xin Ma received the B.E. degree in electronic information engineering from Jiangsu University (JSU), Jiangsu, China, in 2018. He is currently pursuing the M.S. degree in computer technology with the University of Chinese Academy of Sciences (UCAS), Beijing, China. His research interests include image super-resolution, image inpainting, and machine learning.

Zhihang Li received the Ph.D. degree from the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2021. His research interests include deep learning, computer vision, biometrics, and machine learning.

Jie Cao (Member, IEEE) received the B.E. degree in automation from North China Electric Power University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include biometrics, pattern recognition, computer vision, and machine learning.

Ran He (Senior Member, IEEE) received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, Dalian, China, 2001 and 2004, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2009. Since September 2010, he has been with NLPR, where he is currently a Full Professor. His research interests focus on information theoretic learning, pattern recognition, and computer vision. He is an IAPR Fellow. He serves as an Associate Editor for the *Neurocomputing* (Elsevier), and serves on the program committee of several conferences.