IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received December 11, 2019, accepted February 28, 2020, date of publication March 11, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980047

Non-Visual to Visual Translation for Cross-Domain Face Recognition

HAN BYEOL BAE, TAEJAE JEON, YONGJU LEE, SUNGJUN JANG, AND SANGYOUN LEE, (Member, IEEE)
Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea

Corresponding author: Sangyoun Lee (syleee@yonsei.ac.kr)

This work was supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289).

ABSTRACT Reducing the cross-modality gap between two different domains is a challenging problem for heterogeneous face recognition (HFR). The current visual domain face recognition system is not easy to solve the discrepancy of cross-modality when two comparing domains are heterogeneous. Moreover, the amount of HFR dataset is significantly insufficient, making it considerable difficulty in training. This paper proposes a novel two-step framework that consists of the image translation module and the feature learning module to obtain an enhanced cross-modality matching system for heterogeneous datasets. First, the image translation module consists of a Preprocessing Chain (PC) method, CycleGAN, and the Siamese network. It enables to meet the conditions for preserving contents along with changing the styles from the source domain to the target domain. Second, in the feature learning module, the training dataset and its translated images are used together for fine-tuning the pre-trained backbone model in the visual domain. This allows for discriminative and robust feature matching of the probe and gallery test datasets in the visual domain. The experimental results are evaluated with two scenarios, using the CUHK Face Sketch FERET (CUFSF) dataset and the CASIA NIR-VIS 2.0 dataset. The proposed method achieves a better recognition performance in comparison to the state-of-the-art methods.

INDEX TERMS Cross-modality gap, heterogeneous face recognition (HFR), image preprocessing, image-to-image translation, NIR-VIS face matching, sketch-VIS face matching, supervised feature learning.

I. INTRODUCTION

Face recognition (FR) is one of the important research topics in machine learning and pattern recognition. FR has been regarded as a relatively accessible and reliable technology in comparison to other biometric technologies. In recent years, FR research has been further improved with the development of deep learning technique. Deep neural network architectures [1], [2], creating novel loss functions [2]–[4], and large-scale face datasets [5]–[7] are the factors that made it possible to increase performance. Despite these improvements, there are still many challenging tasks in face recognition topic. Conventional face recognition systems identify people by comparing the visual images under homogeneous conditions. However, recent intelligent security and criminal investigation scenarios demand matching cross domains in heterogeneous environments [8]–[10]. This cross-modality gap occurs in various cases, such as various camera sensors, camera resolution differences, and comparing sketch and photo. Therefore conventional face recognition systems generally do not guarantee a high recognition rate for Heterogeneous face recognition (HFR).

In most cases, HFR involves a gallery dataset consisting of visual images (VIS). Probe images can be from any modality, such as near infrared (NIR), thermal infrared (TIR), and sketch images. Many types of research have attempted to address the problem of reducing the modality gap between cross-modality face pairs using handcraft approaches and deep learning-based approaches. Conventionally, handcraft approaches are classified as synthesis-based methods, common subspace projection-based methods, and invariant local feature-based methods. The handcraft approaches show an excellent performance in situations like the well-aligned frontal images along with the small changing conditions of texture and lighting. However, if these conditions are not met, this approach does not provide an optimal solution.

The associate editor coordinating the review of this manuscript and approving it for publication was Ivan Lee.

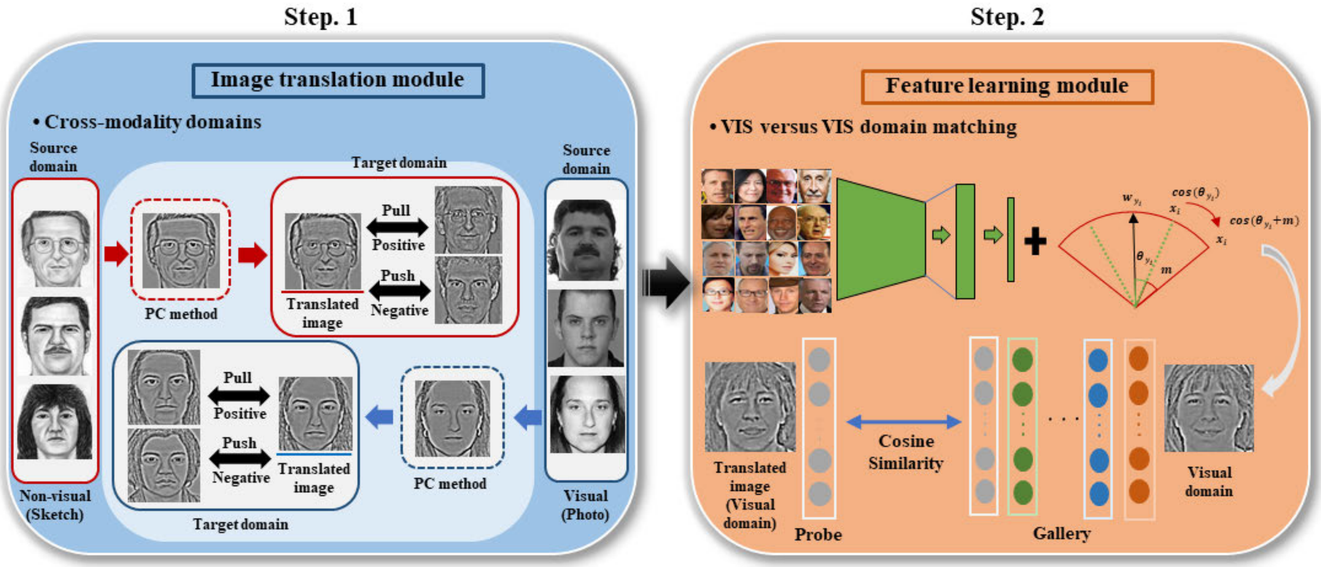

FIGURE 1. Illustration of the flow chart of the proposed framework consisting of the image translation module and the feature learning module. In the first step, the image translation module is used to maintain the contents of the translated image and to find the optimal style change to each target domain for the visual and non-visual images. In the second step, the feature learning module enables extracting the discriminative embedding vector through the backbone network and feature matching of the probe and gallery test datasets in the visual domain.

The recent deep learning-based approach has suggested a discriminative and robust learning method; however, there are limitations when relying on large-scale training data that cannot be applied to HFR datasets with less training data.

This paper proposes a two-step framework consisting of the image translation module and the feature learning module to improve the recognition performance by reducing the cross-modality gap for heterogeneous face recognition. There are two forms of motivation in this framework. First, the image-to-image translation enables the transformation of different domain images in the unpaired setting. Second, when the translated images are obtained from cross different domains, they can be used to provide additional information by fine-tuning the pre-trained model of the visual domain face recognition system. Therefore, we can acquire a discriminative model that can reduce modality gaps between domains. Then, feature matching is performed to between the probe images translated from non-visual to visual and the visual gallery images.

Concretely, the overall of our proposed framework is shown in Fig. 1. Also, a detailed illustration of each module is presented in Fig. 2 and Fig. 5, respectively. For the first step, the image translation module consists of a Preprocessing Chain (PC) method [11], CycleGAN [12], and the Siamese network. The start of the image translation module is applying the PC method to minimize the illumination variations for the HFR dataset. This preprocessing method facilitates image translation by normalizing images of different domains. Then, the preprocessed images in the cross domains are used in image translation to the target domain, respectively. The training process of image translation is carried out by integrating the proposed Siamese network into CycleGAN. The training sequence is carried out as follows. A training image is first used to train the generator of CycleGAN, followed by the discriminator of CycleGAN, and the layers of the Siamese network. And we used contrastive loss [13] in the Siamese network. The basic ideas of contrastive loss are to wide the inter-class distance and to narrow the intra-class distance. Since contrastive loss requires image pairs as inputs, the proposed Siamese network receives the image pairs as follows: In other words, the translated image and the positive sample image are pulled together in the target domain, and the translated image and negative sample are pushed to each other in the target domain.

The feature learning module is the second step of the framework. The same class of translated images and their corresponding target domain images are labeled as the same label. Then, the training dataset and its translated images are used to fine-tune the pre-trained backbone model to obtain discriminative embedding vector. Therefore, this can not only reduce the difference of cross-modality but also carry out feature matching of the probe and gallery test dataset in the visual domain. The experimental results demonstrate that the proposed framework shows a better recognition performance than the state-of-the-art methods in the CUHK Face Sketch FERET (CUFSF) dataset and the CASIA NIR-VIS 2.0 dataset.

In this paper, the contributions of our proposed framework are as follows. First, by applying PC method to HFR datasets, the illumination variations are minimized in cross-domain images. Second, we additionally implement the Siamese network to reduce the gap between the translated image and its corresponding pair in the target domain. The Siamese network is integrated with CycleGAN and is trained simultaneously. Therefore, preprocessed translated images can preserve contents while transforming style more appropriate for target images. Lastly, translated images are used to provide additional information by fine-tuning the pre-trained backbone model. By doing so, we can acquire a discriminative embedding vector and this enables us to carry out feature matching of the probe and gallery test dataset in the visual domain.

The rest of this paper is organized as follows: Section II provides a review of works related to HFR. Section III details the proposed framework. This section starts with introducing the network architecture of the framework, describing how to reduce the gaps of cross-modality and how to improve performance. Section IV describes the experimental setup, datasets, and presents an analysis of the results of the experiment. Finally, the conclusion is described in Section V.

II. RELATED WORK

This section briefly reviews the literature for the following three categories of methods minimizing the cross-modality gap: the common subspace projection based methods, invariant feature descriptor based methods, and the synthesis based methods.

A. COMMON SUBSPACE PROJECTION BASED METHODS

Common subspace projection based methods that belong to this approach aim to learn the mapping function that minimizes the cross-modality discrepancy by projecting cross-modality images into a common subspace as close as possible. Lin and Tang [14] proposed a common discriminant feature extraction method, which is used to extract features from cross-modality images and this features are projected into a common feature space. Yi *et al.* [15] proposed canonical correlation analysis (CCA) for face matching between NIR and VIS images. Later, Li extends this approach in [16]. Regression based methods [17]–[20] are proposed to enable learning of mapping functions that connect cross-modality domains and common spaces. Sharma and Jacobs [21] proposed a method to allow linear mapping of cross-modality images with a common subspace where the mutual covariance is maximized. To demonstrate the four heterogeneous scenarios, Klare and Jain [10] proposed a prototype random subspace (P-RS) method.

B. INVARIANT FEATURE DESCRIPTOR BASED METHODS

Invariant feature descriptor based methods focus on extracting invariant features that are not affected by the discrepancy between cross-modality images. Liao *et al.* [22] proposed a method applying multi-block binary patterns (MB-LBP) after difference-of-Gaussian filtering to match the NIR and VIS face images. Klare *et al.* [23] proposed a local feature-based discriminant analysis (LFDA) framework by extracting scale-invariant feature transform (SIFT) [24] and multiscale local binary pattern (MLBP) [25] feature descriptors as a patch unit from sketch and VIS face images. Zhang *et al.* [26] proposed a feature descriptor based on coupled information-theoretic encoding (CITE). CITE captures discriminative local face structures for effective matching between VIS and sketch images. Galoogahi and Sim [27] proposed a local radon binary pattern (LRBP) that applies the local binary pattern after conducting Radon transform of the VIS face images and the sketch face images. Galoogahi and Sim [28] also proposed histogram of averaged oriented gradients (HAOG) to reduce the discrepancy between the VIS and sketch face images. Recently, Gong *et al.* [29] proposed a common encoding feature discriminant (CEFD) approach to extract discriminative common features by transforming the cross-modality face images into a common encoding space. Roy and Bhattacharjee [30] proposed the local maximum quotient (LMQ) to extract the invariant features in the cross-modality face images. Peng *et al.* [31] proposed a graphical representation-based HFR (G-HFR). In other studies [32]–[35] used convolutional neural network (CNN)-based architecture to find invariant feature space of HFR datasets.

C. SYNTHESIS METHODS

The synthesis method is a method transforming the different modality images into the same modality. Tang and Wang [36], [37] first proposed the method transforming the photo to sketch using the eigenface method. Liu *et al.* [38] proposed a Locally Linear Embedding (LLE) method for transforming the pictures into sketches based on image patches. A series of Markov model-based approaches have been proposed to consider the relationship between the adjacent local patches [39]–[41]. Wang *et al.* [42] proposed the transductive learning method to reduce the high loss in training samples. To reduce the loss of high-frequency information, Gao *et al.* [43] proposed a sparse neighbor selection and spare-representation-based enhancement (SNS-SRE). Later, Wang *et al.* [44] proposed sparse feature selection and supporting vector revision (SFS-SVR). Additionally, Wang *et al.* [45] proposed a quick method to generate a sketch. Peng *et al.* [46] proposed a Markov model-based framework learning the weights of the candidate for multi-representation and target image patches adaptively. Recently, many approaches [12], [47] based on the generative adversarial networks (GANs) proposed by Goodfellow *et al.* [48] made it possible to obtain a more photo-realistic synthesis image than the existing methods. Additionally, other researchers [49], [50] have employed the GANs to generate VIS face images from TIR face images. Song *et al.* [51] proposed domain-invariant feature learning by generating a VIS face image from a NIR face image using GANs. Cao *et al.* [52] used GANs to augment the intra-class data in the proposed framework. The advantage of the synthesis-based approach is that it can apply the conventional visual domain face recognition system. However, this approach shows less detail on non-facial areas and requires a lot of training data. Therefore, the purpose

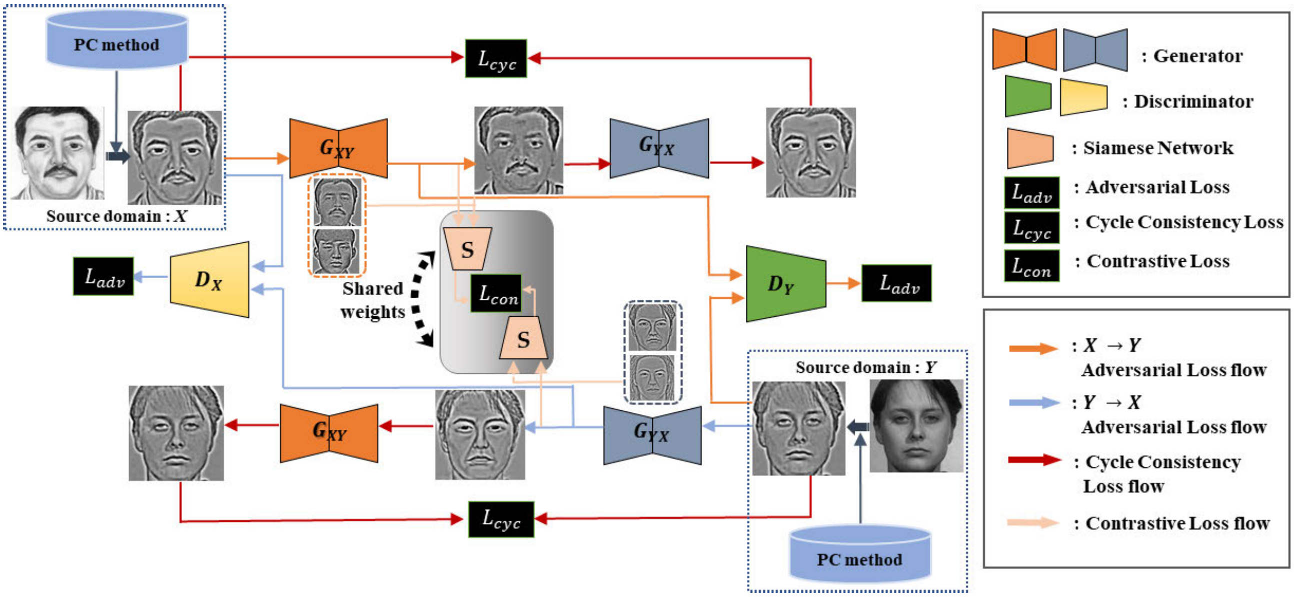

FIGURE 2. Illustration of the proposed image translation module consisting of the Preprocessing Chain (PC) method, CycleGAN, and the Siamese network. CycleGAN learns mapping functions (G_{XY} and G_{YX}) between two domains (X, Y), and the Siamese network simultaneously learns a latent space adding constrains in the learning procedure of mapping functions. The network and loss flows used to learn the network are presented in the illustration.

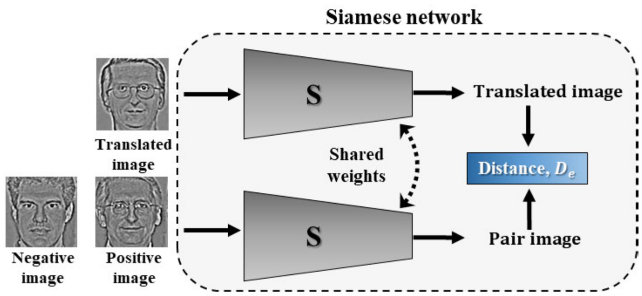

FIGURE 3. Illustration of the proposed Siamese network structure. The weights of each network in the Siamese network are shared by each other.

of the proposed method is to preserve the contents suitable for the target domain while enabling visual-to-visual matching using the existing visual domain face recognition system.

III. PROPOSED METHOD

This section presents the proposed framework to improve the recognition performance by reducing the cross-modality gap in the HFR datasets. We first explain the baseline overview, then the image translation module and the feature learning module.

A. BASELINE OVERVIEW

The purpose of our proposed framework is to reduce cross-modality gap between different domains to perform discriminative and robust feature matching between probe and gallery test dataset in the visual domain as illustrated in Fig. 1. The framework consists of two steps: the image translation module and the feature learning module. Various types of image translation methods [12], [47], [53], [54] that use GANs have recently produced impressive results. The GANs, which consists of a generator and a discriminator network, is used to train the generator to produce the most realistic image in order to prevent the discriminator network from distinguishing between the real and fake images. Meanwhile, the discriminator is trained to distinguish between the real and fake images. This paper adopted the unpaired uni-modal image translation to overcome the constraints due to the lack of paired data and to transform the images into the effective target domain. CycleGAN [12] is used as the baseline for image translation, which is the first step of the framework as illustrated in Fig. 2. Zhu *et al.* proposed CycleGAN by adopting a cycle consistency loss for the translated image. The translated image should recover to the original image after a cycle of translation and reverse translation. Therefore, CycleGAN needs two generator-discriminator pairs. Since CycleGAN learns mapping functions by separating the latent space of the two generators, the translated image makes it possible to follow a specific style in the target domain. However, if the structural variation between the source domain and the target domain is significant, there is no guarantee that the translated image preserves the contents of the input image. For example, when looking at the HFR datasets such as the VIS-NIR and VIS-sketch, it is not easy to preserve the contents if the image is highly exaggerated or if the spectrum range of the camera sensor is too different. Therefore, the Siamese network is added to the image translation module to preserve the contents of the translated image in the target domain as illustrated in Fig. 2. The basic idea of the Siamese network was first proposed to solve the verification problem in [55], [56]. The Siamese network extracts 128-dimensional

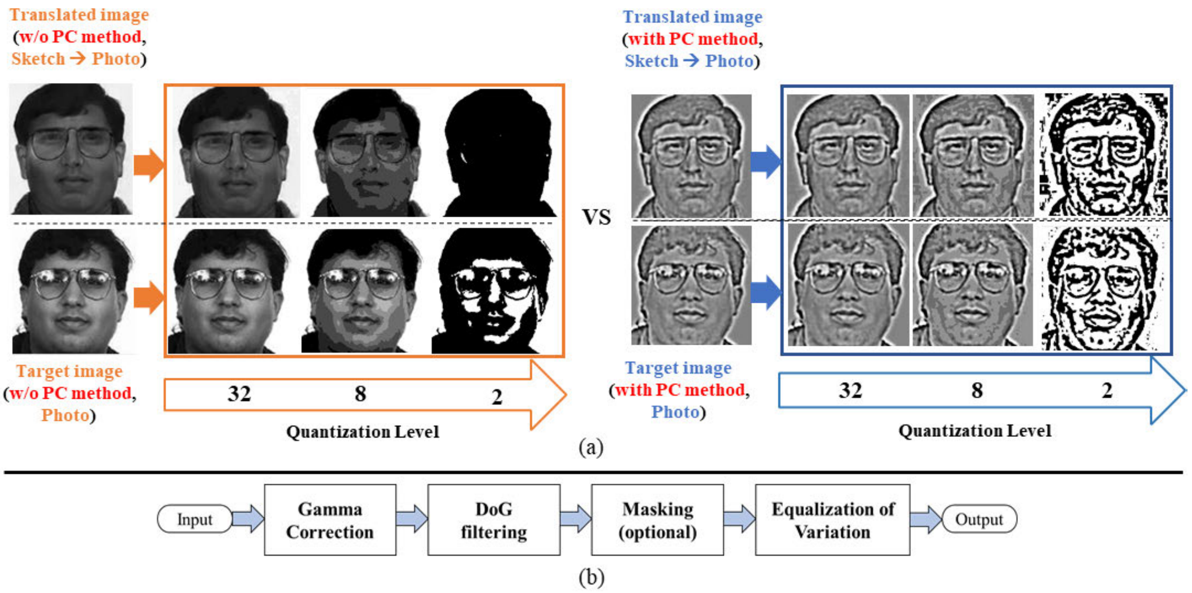

FIGURE 4. (a) The quantization level results according to the presence or absence of the Preprocessing Chain (PC) method in the translated image and its corresponding target image. (b) The stages of the PC method; gamma correction, Difference of Gaussian (DoG) filtering, masking, and equalization of variation.

embedding vector through the CNN-based network and uses contrastive loss [13]. Positive and negative pairs are required to calculate the contrastive loss of the translated image. To satisfy these conditions, we place the translated image of the target domain and their positive and negative pairs in the Siamese network inputs. Also, the weights of each network in the Siamese network are shared with each other, as shown in Fig. 3. CycleGAN and the Siamese network are trained together in the following order: the training dataset is first used to learn the generators, then the discriminators, and finally, the layers in Siamese. However, the translated image cannot guarantee that the range of intensity is similar between the translated image and the target domain images. Therefore, the Preprocessing Chain (PC) method proposed by Tan and Triggs [11] is also applied to the dataset before image translation is carried out. Results depending on quantization level and stages of the image preprocessing method are depicted in Fig. 4. When comparing the result images according to the quantization level, the intensity level difference between the target image and translated image is much smaller in the result of applying PC method. After the image translation module step, the same class of translated images and their corresponding target domain images are labeled as the same label. The second step, the feature learning module, is then performed as demonstrated in Fig. 5. The residual networks [57] are chosen as the CNN-based backbone model. After backbone network is pre-trained on the large-scale dataset [7], the recombined training dataset are used to fine-tune the pre-trained backbone model. This allows the acquisition of discriminative and robust embedding vector to enable feature matching of probe and gallery test datasets in the visual domain.

TABLE 1. The generator, discriminator architecture of CycleGAN [12].

Network	Generator
1	Conv(N64, K7, S1) - ReLU
2	Conv(N128, K3, S2) - InstanceNorm - ReLU
3	Conv(N256, K3, S2) - InstanceNorm - ReLU
4	Residual Block(N256)
5	Residual Block(N256)
6	Residual Block(N256) - Dropout(0.5)
7	Residual Block(N256) - Dropout(0.5)
8	Residual Block(N256) - Dropout(0.5)
9	Residual Block(N256) - Dropout(0.5)
10	Residual Block(N256) - Dropout(0.5)
11	Residual Block(N256) - Dropout(0.5)
12	Residual Block(N256) - Dropout(0.5)
13	Dconv(N128, K3, S2) - InstanceNorm - ReLU
14	Dconv(N64, K3, S2) - InstanceNorm - ReLU
15	Conv(N1, K7, S1) - Tanh
Network	**Discriminator**
1	Conv(N64, K4, S2) - LeakyReLU(0.2)
2	Conv(N128, K4, S2) - InstanceNorm - LReLU(0.2)
3	Conv(N256, K4, S2) - InstanceNorm - LReLU(0.2)
4	Conv(N512, K4, S2) - InstanceNorm - LReLU(0.2)
5	Conv(N1,K4,S1) - Sigmoid

The Conv(Nα, Kβ, Sγ) is denoted as a convolution layer with α filters, β x β kernel, and γ stride. The Dropout(0.5) means droptout rate is 0.5. The Residual Block(Nα) is denoted as contains 3 x 3 convolutional layers with the α filters on both layer. The DConv(Nα, Kβ, Sγ) is denoted as deconvolution layer with α filters, β x β kernel, and 1/γ stride. LReLU is the LeakyReLU. In addition, LReLU(0.2) means LReLU with slope 0.2.

B. IMAGE TRANSLATION MODULE

1) NOTATION AND NETWORK ARCHITECTURE

Image translation is performed between different domains as the first step to reduce the modality gap in the cross-domain, as illustrated in Fig. 1 and Fig. 2. The goal of this

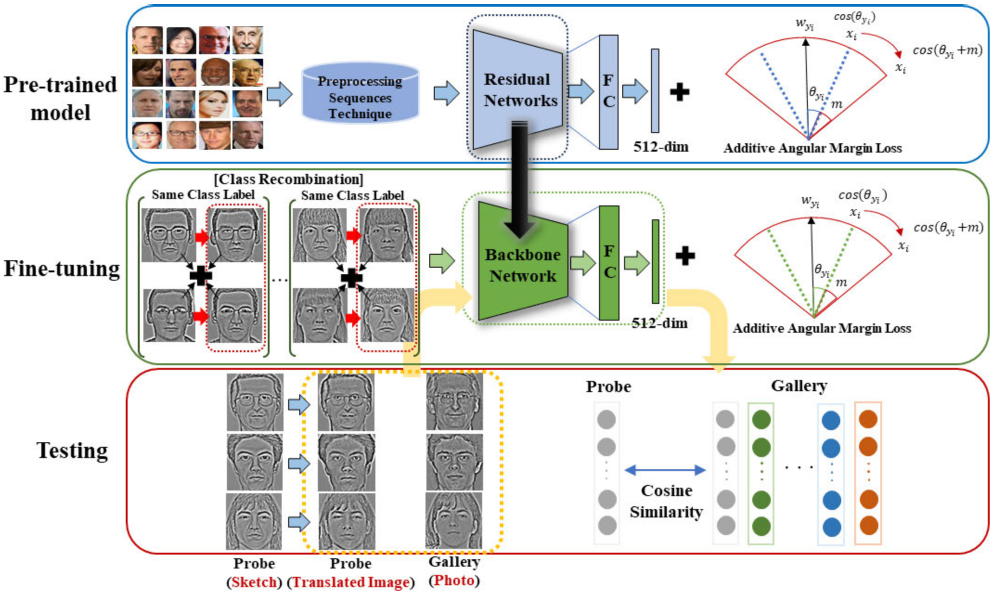

FIGURE 5. **Illustration of the proposed feature learning module. A CNN-based residual networks is selected and this is pre-trained with a large-scale dataset to carry out supervised feature learning. The same class of translated images and their corresponding target domain images are labeled as the same label and used as fine-tuning data. Feature matching between the probe and gallery images in the visual domain is performed with this fine-tuned backbone model.**

step is to learn mapping functions between two domains, X and Y, when given training samples $x_i \in X$ and $y_i \in Y$. The data distribution is denoted as $x \sim p_{data(x)}$ and $y \sim p_{data(y)}$. To transform images to the style of the target domain, two generator-discriminator $\{G_{XY}, D_Y\}$ and $\{G_{YX}, D_X\}$ are required for two domains: X and Y. The $G_{XY} : X \rightarrow Y$ indicates mapping from the X domain to the Y domain, and $G_{YX} : Y \rightarrow X$ indicates mapping from the Y domain to the X domain. Additionally, the adversarial discriminators $\{D_X, D_Y\}$ distinguish whether the inputs are real images or fake images. The network architecture is same as CycleGAN [12], and its details are represented in Table 1. The generators consist of the convolution layers, nine residual blocks, and the deconvolution layers. Moreover, PatchGAN [47] is used for the discriminators. CycleGAN can generate the translated images to the style of the target domain because they are separated from each other without sharing latent space in each translation process. Additionally, since there are no ground truth images for the translated images, the guide for maintaining the contents of the target domain does not exist. This drawback is solved by the CycleGAN through the cycle consistency loss. In this process, the translated image should be recovered to original image after a cycle of translation and reverse translation. However, when the discrepancy of the cross-modality is large, like the HFR dataset, the cycle consistency loss is not enough to maintain the contents of original image. Thus, the additonal siamese network is integrated with CycleGAN and both are trained simultaneously. The goal of the Siamese network is to learn a general similarity function. By measuring and comparing the similarity of the embedding vectors between the translated image and their corresponding pair in the target domain, this network preserves the contents of the translated image in the target domain. The architecture details of a CNN-based Siamese network are represented in Table 2. Each translated image and their corresponding pair go through a network consisting of two branches during training as presented in Fig. 3. The outputs of these branches are used for optimization using a contrastive loss [13]. The loss tries to minimize the squared Euclidean distance between the embedding vectors of positive image pairs and to maximize that of the negative pairs. However, the difference from the conventional contrastive loss is that the conventional one does not check with the binary labels whether the pair is positive or negative. Data was inputted along with their corresponding positive and negative pair to supply the consistency through the contrastive loss. In other words, the contrastive loss L_{con} pulls a translated image and its positive pair in the target together, and push the translated image and negative pair in the target apart, according to the following equations:

$$\begin{aligned} L_{con}(x_1, x_2, x_3) &= L_{con}^{negative\ pair}(x_1, x_2) \\ &\quad + L_{con}^{positive\ pair}(x_1, x_3), \\ L_{con}^{negative\ pair}(x_1, x_2) &= \{max(0, m - D_e(x_1, x_2))\}^2, \\ L_{con}^{positive\ pair}(x_1, x_3) &= D_e(x_1, x_3)^2. \end{aligned} \tag{1}$$

where x_1 is the embedding vector of input image. x_2 and x_3 are embedding vectors of corresponding pair images of input image. D_e denotes the Euclidean distance between two embedding vectors. m is the margin that defines the separability in the embedding space. Here, the Euclidean distance metric between two embedding vector is defined, according

TABLE 2. The Siamese network architecture.

Network	Siamese network
1	Conv(N64, K4, S2) - ReLU - MaxPooling
2	Conv(N128, K4, S2) - ReLU - MaxPooling
3	Conv(N256, K4, S2) - ReLU - MaxPooling
4	Conv(N512, K4, S2) - L2Normalization
5	FC(128)

The Conv(Nα, Kβ, Sγ) is denoted as a convolution layer with α filters, β x β kernel, and γ stride. FC is fully-connected layer. In addition, FC(128) means the 128-dimensional output.

to the following equation:

$$D_e(S(G_{XY}(x)), S(y_{xp})) = \|S(G_{XY}(x)) - S(y_{xp})\|_2. \quad (2)$$

where y_{xp} is defined as the x corresponding positive pair in the Y domain and y_{xn} is the x corresponding to the negative pair in Y domain. S is the Siamese network.

2) LOSS FUNCTION

As illustrated in Fig. 2, the image translation module is optimized by three objectives: the adversarial loss, the cycle consistency loss, and the contrastive loss. Two generator-discriminator pairs, $\{G_{XY}, D_Y\}$ and $\{G_{YX}, D_X\}$, are required to enable the image translation of the unpaired cross-domain images. Using the adversarial loss [48], the generators learn the style of the target domain to prevent the discriminators from distinguishing the images generated. In contrast, the discriminators learn to classify the images in their each domain. The adversarial loss can be expressed according to the following equation:

$$L_{adv}(G_{XY}, G_{YX}, D_Y, D_X) = L_{Yadv}(G_{XY}, D_Y) + L_{Xadv}(G_{YX}, D_X). \quad (3)$$

For generator G_{XY} and its corresponding discriminator D_Y, the adversarial loss is expressed with the following equation:

$$L_{Yadv}(G_{XY}, D_Y) = \mathbb{E}_{y\sim p_{data(y)}}[logD_Y(y)] + \mathbb{E}_{x\sim p_{data(x)}}[log(1 - D_Y(G_{XY}(x)))]. \quad (4)$$

For the generator G_{YX} and its corresponding discriminator D_X, the adversarial loss is determined by the following equation:

$$L_{Xadv}(G_{YX}, D_X) = \mathbb{E}_{x\sim p_{data(x)}}[logD_X(x)] + \mathbb{E}_{y\sim p_{data(y)}}[log(1 - D_X(G_{YX}(y)))]. \quad (5)$$

For the minmax optimizaiton, $\{G_{XY}, G_{YX}\}$ aim to minimize this objective against $\{D_X, D_Y\}$, while $\{D_X, D_Y\}$ tries to maximize it, i.e.,

$$G^*_{XY}, G^*_{YX}, D^*_Y, D^*_X = \arg\min_G \max_D L_{adv}(G_{XY}, G_{YX}, D_Y, D_X). \quad (6)$$

However, the above general adversarial GAN loss need some changes. Since general adversarial GAN loss is a form of cross-entropy, the valuable gradient feedback may not be delivered to the generator. Therefore, to stabilize the training procedure, we replace Eq.(4) and Eq.(5) with Least Square GAN objective [58], as the following equations:

$$\begin{aligned} L_{Yadv}(G_{XY}, D_Y) &= \mathbb{E}_{y\sim p_{data(y)}}[(D_Y(y) - 1)^2] \\ &\quad + \mathbb{E}_{x\sim p_{data(x)}}[D_Y(G_{XY}(x))^2], \quad (7) \\ L_{Xadv}(G_{YX}, D_X) &= \mathbb{E}_{x\sim p_{data(x)}}[(D_X(x) - 1)^2] \\ &\quad + \mathbb{E}_{y\sim p_{data(y)}}[D_X(G_{YX}(y))^2]. \quad (8) \end{aligned}$$

Next, in the loss function L_{cyc} [12], the L_1 norm is used in the following equation:

$$L_{cyc}(G_{XY}, G_{YX}) = \mathbb{E}_{x\sim p_{data(x)}}[\|G_{YX}(G_{XY}(x)) - x\|_1] + \mathbb{E}_{y\sim p_{data(y)}}[\|G_{XY}(G_{YX}(y)) - y\|_1]. \quad (9)$$

Adding this cycle consistency loss, the range of possible mapping functions gets reduced and it prevents the network from falling into the mode collapse state. As mentioned above, we integrate CNN-based Siamese network with the CycleGAN. A similarity comparison is made between the embedding vectors using the contrastive loss in Eq.(1). A more effective and intuitive approach is used by always exploiting both positive and negative pairs. In this regard, Eq.(1) is changed as the following equation:

$$\begin{aligned} &L_{con}(G_{XY}, G_{YX}, S) \\ &\quad = L^{negative}_{con}(G_{XY}, G_{YX}, S) + L^{positive}_{con}(G_{XY}, G_{YX}, S), \\ &L^{negative}_{con}(G_{XY}, G_{YX}, S) \\ &\quad = \mathbb{E}_{x\sim p_{data(x)}}[max(0, m - D_e(S(G_{XY}(x)), S(y_{xn})))^2] \\ &\qquad + \mathbb{E}_{y\sim p_{data(y)}}[max(0, m - D_e(S(G_{YX}(y)), S(x_{yn})))^2], \\ &L^{positive}_{con}(G_{XY}, G_{YX}, S) \\ &\quad = \mathbb{E}_{x\sim p_{data(x)}}[D_e(S(G_{XY}(x)), S(y_{xp}))] \\ &\qquad + \mathbb{E}_{y\sim p_{data(y)}}[D_e(S(G_{YX}(y)), S(x_{yp}))]. \quad (10) \end{aligned}$$

where y_{xp} is defined as the x corresponding positive pair in the Y domain and y_{xn} is the x corresponding to the negative pair in Y domain. x_{yp} is defined as the y corresponding positive pair in the X domain and x_{yn} is the y corresponding to the negative pair in X domain. S is the Siamese network. Finally, the full objective of image translation module can be defined, which consists of three objectives and can be expressed as follows:

$$L_{Total}(G_{XY}, G_{YX}, D_Y, D_X) = L_{adv}(G_{XY}, G_{YX}, D_Y, D_X) + \lambda L_{cyc}(G_{XY}, G_{YX}) + \gamma L_{con}(G_{XY}, G_{YX}, S), \quad (11)$$

where λ and γ are the weights that control the three objectives.

C. FEATURE LEARNING MODULE

The feature learning module is the second step of the framework as depicted in Fig. 1 and Fig. 5. The feature learning module enables feature matching in the visual domain. After the image translation process, the first step of framework, the translated images are generated in each target domain.

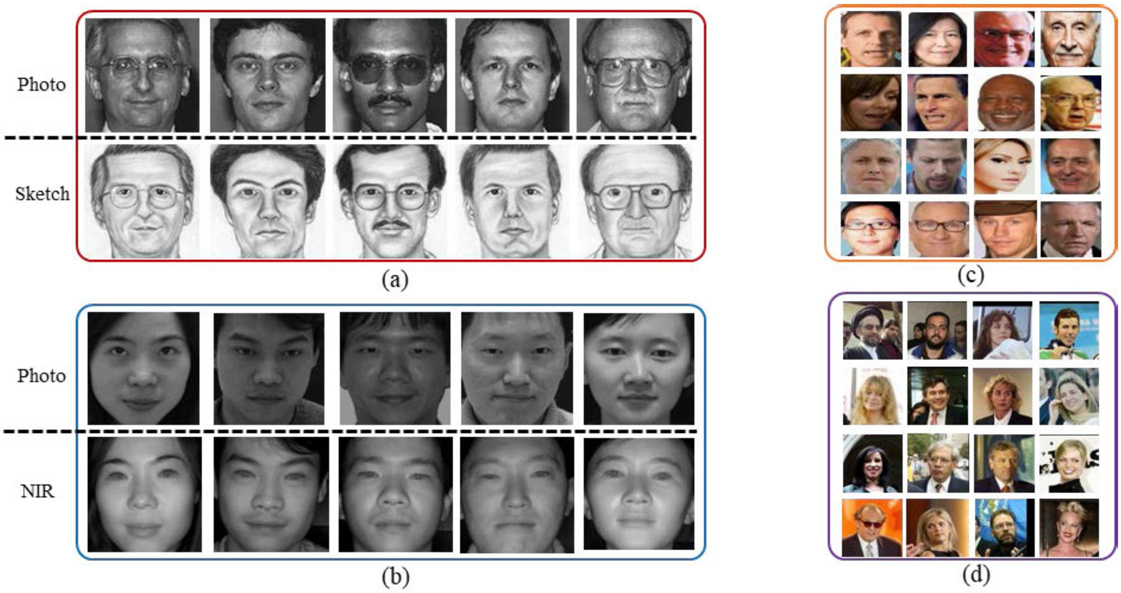

FIGURE 6. Some sample images in datasets. (a) CHUK Face Sketch FERET (CUFSF), (b) CASIA NIR-VIS 2.0, (c) Cleaned MS-Celeb-1M, (d) Labeled Faces in the Wild (LFW).

The same class of translated images and their corresponding target domain images are labeled as the same label. We employ the ResNet-101 [57] as the pre-trained backbone model. We make pre-trained models with cleaned Celeb-1M dataset [7] as backbone model. Then, the recombined training dataset is used to fine-tune the pre-trained backbone model for obtaining discriminative embedding vector. To get the best backbone model, we use the additional angle margin loss (ArcFace) proposed by Deng *et al.* [4].

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{scos\theta_j}}, \tag{12}$$

where N and n are the batch size and the class number, respectively. θ_{y_i} is the target (groud truth) angle. m is the angular margin penalty and s is the feature scale.

As such, we obtain the backbone network that extracts 512-dimensional embedding vector. Therefore, we can perform discriminative and robust feature matching of probe and gallery test dataset in the visual domain.

IV. EXPERIMENTS

To verify and evaluate the proposed method, two datasets are adopted. These datasets are the CUHK Face Sketch FERET (CUFSF) Dataset [26] and the CASIA NIR-VIS 2.0 Face Dataset [59]. Both datasets are the most general and widely used datasets that are open to public. For this paper, the cleaned MS-Celeb-1M dataset [7] is used as a training dataset in second step of the framework to create the backbone model. Additionally, the Labeled Faces in the Wild (LFW) dataset [5] is used as validation dataset to validate these models.

A. DATASETS

1) CUHK FACE SKETCH FERET (CUFSF) DATASET

CUHK Face Sketch FERET (CUFSF) Dataset [26] is widely used among viewed sketch datasets. There are 1,194 subjects in the FERET dataset [60] and every image is a frontal face. Photos have illumination variations and exaggerated sketches were drawn along with the photos as illustrated in Fig. 6(a).

2) CASIA NIR-VIS 2.0 DATASET

The CASIA NIR-VIS 2.0 Face dataset [59] is the largest and most challenging NIR-VIS dataset due to the large variations in lighting, expression, and pose as shown in Fig. 6(b). The CASIA NIR-VIS 2.0 consists of 5,093 VIS images and 12,487 NIR images. There are four sessions with 725 identities, each with 1 to 22 VIS and 5 to 50 NIR images. And this is organized for a 10-fold of experiments. For the training set, there are about 2,500 VIS and 6,100 NIR images from 360 identities. For the test set, the probe set consists of about 6,000 NIR images from 358 identities, and the gallery set consists of only one VIS image from 358 identities.

3) CLEANED MS-CELEB-1M AND LABELED FACES IN THE WILD (LFW) DATASET

The cleaned MS-Celeb-1M dataset proposed by Xu *et al.* [7] is used as a training dataset in the second step of the framework to create the backbone model. The performance is tested with the Labeled Faces in the Wild (LFW) dataset [5], which is used as a validation set. Guo *et al.* [6] produced the MS-Celeb-1M dataset as one of the most popular datasets in large-scale datasets. Guo *et al.* used one million celebrities for the dataset and released 99,892 celebrities for the original training dataset; however, the released dataset contains a lot of noise. For example, some images labeled as one celebrity are actually belonged to other celebrities. Some images are blur and others do not contain human faces. Additionally, the distribution of the original training data is unbalanced. Therefore, to compensate for these problems and to maximize the efficiency of the dataset, Xu *et al.* refined the MS-Celeb-1M dataset. The MS-Celeb-1M dataset refined by Xu *et al.* consists of 100K classes and 5,084,127 images.

The LFW dataset is one of the most popular benchmark dataset. The LFW dataset includes 13,233 face images from 5,749 different identities and provides 6,000 face pairs for the verification protocol under unrestricted conditions. Fig. 6(c) and Fig. 6(d) show examples of the cleaned MS-Celeb-1M and LFW.

B. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

In experiments, our networks were implemented using TensorFlow and PyTorch. The experiments were carried out on a desktop computer with Intel(R) Core(TM) i7 CPU @ 3.20 GHz and 16.0GB RAM. And all of the networks in this paper were learned using NVIDIA GTX1080-TI GPU. Before performing this method, all images of the datasets were cropped as a 128 x 128 size using the multitask cascaded convolutional networks (MTCNN) detector [61]. For images that cannot be processed by MTCNN, we manually cropped those images based on the position of eyes, nose, and mouth. These cropped images were then subsequently normalized using Preprocessing Chain (PC)[1] method [11] to reduce illumination variations.

In the first step, the image translation module were trained based on the following fixed parameter settings in Eq.(11): $\lambda = 10$, $\gamma = 2$, $m = 2$. The network weights of each layer were initialized by a Gaussian distribution with a zero mean and a standard deviation of 0.001. For optimization, Adam [62] was employed where $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the batch size was set to 1. The initial learning rate was 0.0002, and was maintained for the first 100 epochs; however, it linearly decayed to zero over the next 100 epochs. To satisfy the requirement of the image size of the image translation module, the input images were resized to 256 x 256. The embedding size used for the Siamese network was set to the 128-dimension.

In the second step, feature learning module, after the cleaned MS-Celeb-1M and LFW dataset are used to make the pre-trained backbone model, HFR training dataset were used to fine-tune the pre-trained backbone model, and all images of datasets were resized to 112 x 112. The ResNet-101 [57] is employed as the pre-trained backbone model. Additionally, the backbone network was integrated with the ArcFace [4] as a classifier network to find the best model. The LFW dataset was also used as a validation set for selecting the optimal model. The model obtained an accuracy of 99.3% on the LFW dataset. The hyper-parameters of the ArcFace, s and m, were set to 30 and 0.5, respectively, in Eq.(12). The network weights of each layer were initialized by a Gaussian distribution with a zero mean and a standard deviation of 0.001. For optimization, the stochastic gradient descent was employed where the momentum was 0.9, the weight decay was 0.0005, and the batch size was set to 128. The learning rate initially set as 0.001 and maintained for the first 50 epochs,. Afterwards, the step decayed, i.e., the learning rate was multiplied by 0.1 over the next 50 epochs. The generated backbone model in the network was used in the same structure, when fine-tuned with the ArcFace. The embedding size used for the feature matching was set to the 512-dimension.

TABLE 3. Confusion matrix.

		Predict	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

2) EVALUATION METHODS

We divided the HFR database and used it in the experiment for training data and test data. For CUFSF dataset, we divided 500 subjects as training dataset and the remaining 694 subjects as test dataset. For CASIA NIR-VIS 2.0 dataset, we followed View 2 evaluation protocol, which consists of sub experiments. For the performance test of the proposed method, all of the experiments were repeated ten times, and their average was taken as a result of this experiment.

The evaluation of translated image was conducted as a qualitative evaluation because the translated image has no ground truth image in target domain. The evaluation of recognition performance is conducted by the identification rate using cosine similarity. The cosine similarity function was defined as shown in Eq.(13). Also, the additional evaluation of recognition performance is conducted by the verification rate of the specific false acceptance rate (FAR) in Eq.(14) and is calculated by using the confusion matrix as demonstrated in Table 3.

$$Cosine\ Similarity\ (A, B) = \frac{A \cdot B}{\|A\|\|B\|}. \tag{13}$$

$$FAR\ (False\ Acceptance\ Rate) = \frac{FP}{FP + TN}. \tag{14}$$

C. RESULTS

1) QUALITATIVE EVALUATION

The translated images of the visual and non-visual domain obtained from the first step, image translation module, are used as important information for reducing cross-modality gaps in the second step, feature learning module. Therefore, the translated image must not only conform to the style of the target domain but also maintain its contents.

[1] Source code is available at http://lear.inrialpes.fr/people/triggs/src/amfg07- demo- v1.tar.gz..

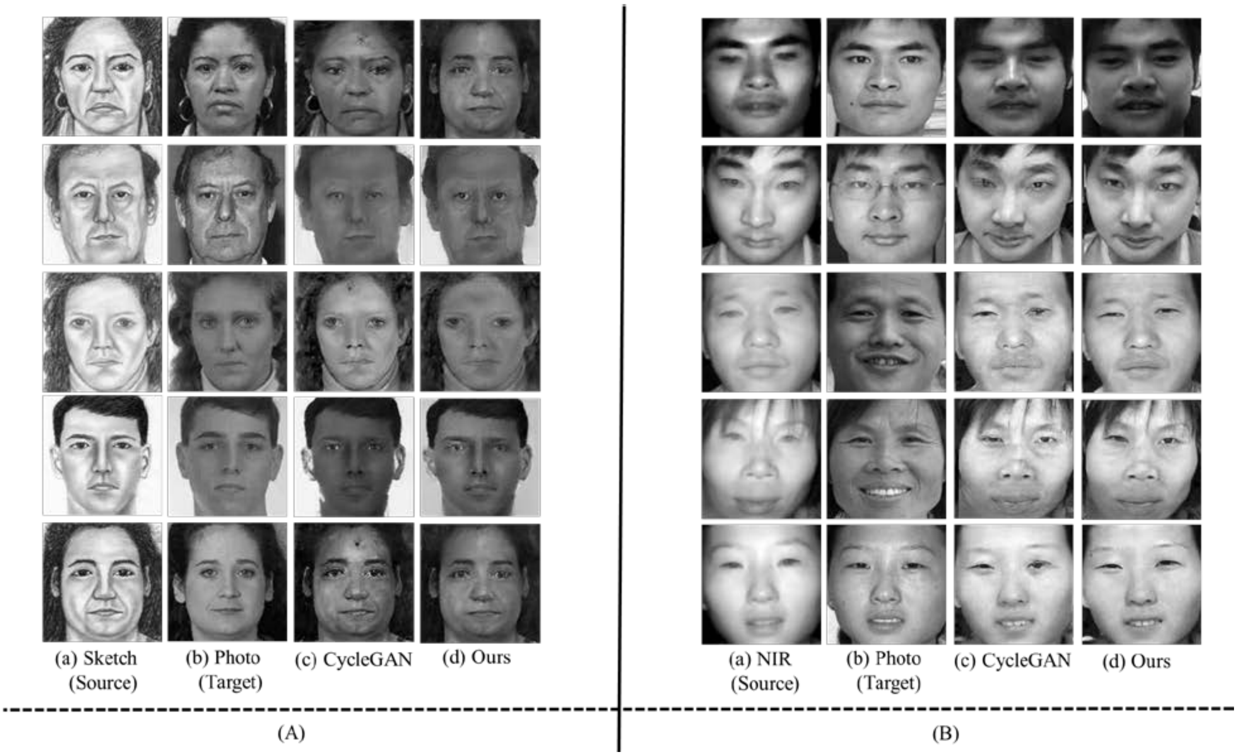

FIGURE 7. Qualitative evaluations of the image translation modules in (A) CUHK Face Sketch FERET (CUFSF) dataset and (B) CASIA NIR-VIS 2.0 dataset, respectively.

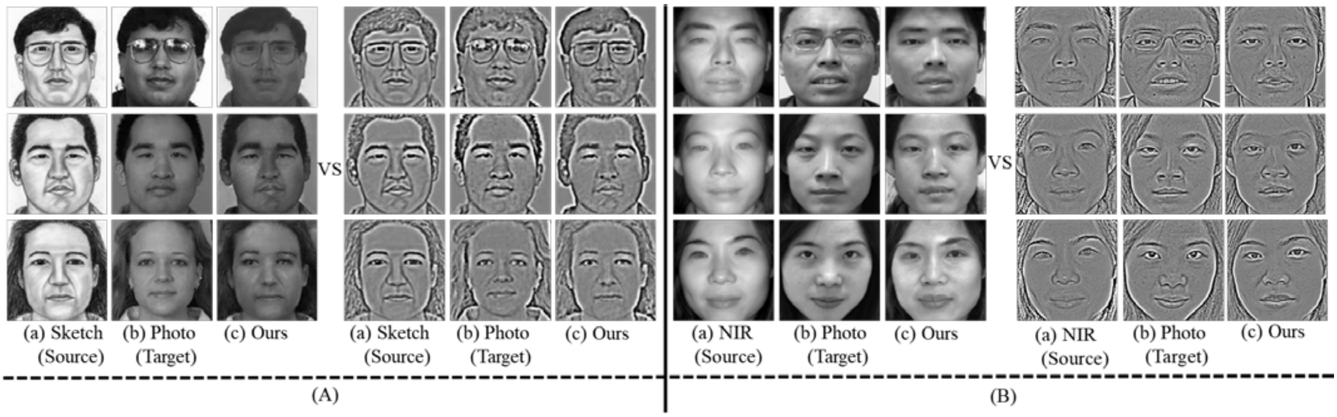

FIGURE 8. Comparison of qualitative evaluations according to whether Preprocessing Chain (PC) is applied or not for image translation modules in (A) CUHK Face Sketch FERET (CUFSF) dataset and (B) CASIA NIR-VIS 2.0 dataset, respectively.

However, the evaluation methods for the Peak-Signal-to-Noise Ratio (PSNR) and the Structural Similarity (SSIM) [63] are not available since the ground truth of the translated image does not exist. Thus, a qualitative evaluation of the translated images was conducted. Because most HFR datasets have a large discrepancy between visual and non-visual domain, cycle consistency loss [12] is not sufficient to maintain contents suitable for the target domain. Thus, to make translated image maintain contents stable, we have proposed a method to integrate a CNN-based Siamese network and CycleGAN and train both simultaneously. As shown in Fig. 7, we can observe that the system integrating CycleGAN and the Siamese network maintains the contents of the translated image better than CycleGAN. Because the translated image cannot guarantee the similarity of the range of intensity between the translated image and the target domain image, the preprocessed [11] dataset was applied to the image translation module as shown in Fig. 4. As presented in the results of the Fig. 8, the translated image shows little intensity difference in the target domain, as well as improved contents retention and style changes to the target domain.

TABLE 4. The Rank-1 Accuracy on the CUFSF dataset.

Method	Rank-1 Accuracy
LRBP [27]	91.12%
G-HFR [31]	96.04%
PLS [21]	51.00%
MRF [40]	46.03%
MWF [41]	74.15%
TFSPS [42]	75.94%
RSLCR [45]	72.62%
MrFSPS [46]	73.36%
VGG-16 [64]	48.80%
VGG-19 [64]	50.80%
Inception-ResNet-v2 [65]	56.30%
ResNet-50 [57]	58.30%
ResNet-101 [57]	63.20%
SE-ResNet-50 [66]	57.80%
SE-ResNet-101 [66]	62.30%
CBAM-ResNet-50 [67]	58.20%
CBAM-ResNet-101 [67]	61.80%
Ours (w/o PC method)	97.26%
Ours	98.70%

TABLE 5. The Rank-1 Accuracy and verification rate of 0.1% FAR on the CASIA NIR-VIS 2.0 dataset.

Method	Rank-1 Accuracy	VR@FAR=0.1%
TRIVET [32]	95.70%	91.00%
IDR [33]	97.33%	95.73%
ADFL [51]	98.15%	97.21%
CDL [34]	98.62%	98.32%
W-CNN [35]	98.70%	98.40%
VGG-16 [64]	55.30%	37.70%
VGG-19 [64]	58.50%	41.35%
Inception-ResNet-v2 [65]	60.30%	44.30%
ResNet-50 [57]	64.10%	55.60%
ResNet-101 [57]	65.80%	62.10%
SE-ResNet-50 [66]	63.30%	53.50%
SE-ResNet-101 [66]	64.20%	60.60%
CBAM-ResNet-50 [67]	63.10%	52.60%
CBAM-ResNet-101 [67]	63.80%	58.40%
Ours (w/o PC method)	99.07%	98.67%
Ours	99.40%	98.74%

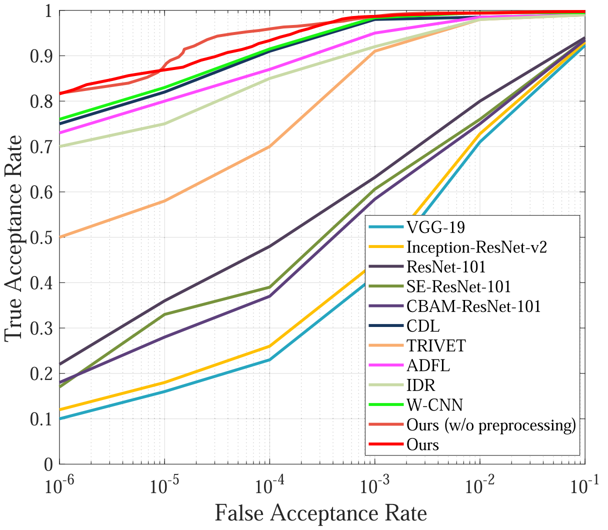

FIGURE 9. The receiver operating characteristic (ROC) curves of the different methods in the CASIA NIR-VIS 2.0 dataset.

2) QUANTITATIVE EVALUATION

A quantitative evaluation was performed in the feature learning module, the second step of the framework. The proposed method was compared with different state-of-the-art methods based on the evaluation metrics of Eq.(13) and Eq.(14).

First, Table 4 shows the rank-1 accuracy on the CUFSF dataset. LRBP [27] and G-HFR [31] are methods based on the invariant feature descriptor. The G-HFR method shows the highest rank-1 accuracy among the methods being compared. The PLS [21] method is one of the common subspace projection based methods; the performance of the PLS method is not adequate. MRF [40], MWF [41], TFSPS [42], RSLCR [45], and MrFSPS [46] methods are synthesis based methods. Overall, the methods have less than 80% performance, and the MRF method has the lowest performance among these methods. We used the several deep learning networks [57], [64]–[67] pre-trained in the visual domain as a comparison method. The results of the experiment show that the performance is poor because the general face recognition system on the visual domain does not extract any discriminative features that can distinguish the images with cross-modality. The proposed method shows the best performance compared with state-of-the-art methods. We also present the performance with and without preprocessing applied to the dataset. The result of applying preprocessing shows 2.66% higher performance than the G-HFR method.

Second, Table 5 shows the rank-1 accuracy and verification rate of 0.1% FAR on the CASIA NIR-VIS 2.0 dataset. We used the recently proposed TRIVET [32], IDR [33], ADFL [51], CDL [34], W-CNN [35], and the saveral deep learning pre-trained models [57], [64]–[67] for comparing the performance with the proposed method. Among compared methods, the TRIVET method shows the lowest performance in both the Rank-1 accuracy and VR@FAR=0.1% performance, whereas W-CNN shows the highest performance. The method proposed in this paper shows the highest performance compared with the state-of-the-art method. When compared with the W-CNN, the rank-1 accuracy was 0.7% higher, and the VR@FAR=0.1% was 0.34% higher. We plot the receiver operating characteristic (ROC) curves of the proposed method and its competitive state-of-the-art methods in Fig. 9. In order to better show the results of analysis on the ROC curve, a semi logarithmic coordinate is used to show the curves. In ROC curves, the proposed method performs significantly better compared with the other methods. In the section where FAR is higher than 1%, all the methods are not significantly different in their verification rates, except for the several methods [57], [64]–[67], which are the pre-trained deep learning models. Applying the HFR dataset directly to the pre-trained models in the visual domain shows low performance. The reason is that it doesn't extract discriminative features in the images with cross-modality. As shown in Fig. 9, it can be seen that the method to which preprocessing is applied has a higher verification rates than the method to which preprocessing is not applied, except for a section in which FAR is more than 0.001% and less than 0.03%.

V. CONCLUSION

This paper proposes a novel two-step framework that consists of the image translation module and the feature learning module. The purpose of the proposed method is to obtain an enhanced cross-modality matching system in the visual domain system. To make this possible, first, we integrate the Siamese network with CycleGAN and train it with a preprocessed HFR dataset. By doing so, the translated images better maintain their contents, while at the same time transforming style more similar to the target domain. Second, the images of the training dataset and its translated images are used to fine-tune the pre-trained backbone model to obtain a discriminative embedding vector. This enables feature matching of probe and gallery test datasets in the visual domain. Overall, the experimental results show that the proposed method performs better than other state-of-the-art methods. However, since our framework can be affected by the amount of training dataset, we plan to consider the unpaired multi-modal image-to-image translation as a method to overcome the limitations on the amount of dataset in the cross-domains in future work.

REFERENCES

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

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[3] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[5] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07–49, Oct. 2007.

[6] 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.* Springer, 2016, pp. 87–102.

[7] Y. Xu, Y. Cheng, J. Zhao, Z. Wang, L. Xiong, K. Jayashree, H. Tamura, T. Kagaya, S. Pranata, S. Shen, J. Feng, and J. Xing, "High performance large scale face recognition with multi-cognition softmax and feature retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1898–1906.

[8] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello, "Automated border control e-gates and facial recognition systems," *Comput. Secur.*, vol. 62, pp. 49–72, Sep. 2016.

[9] S. Klum, H. Han, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. composite sketches," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[10] 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.

[11] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[13] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[14] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2006, pp. 13–26.

[15] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *Proc. Int. Conf. Biometrics*. Springer, 2007, pp. 523–530.

[16] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 605–611.

[17] 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.

[18] Z. Lei, C. Zhou, D. Yi, A. K. Jain, and S. Z. Li, "An improved coupled spectral regression for heterogeneous face recognition," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 7–12.

[19] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1707–1716, Dec. 2012.

[20] 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.

[21] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. CVPR*, Jun. 2011, pp. 593–600.

[22] 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*. Springer, 2009, pp. 209–218.

[23] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[26] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.

[27] H. Kiani Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1837–1840.

[28] H. K. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 224–229.

[29] 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.

[30] H. Roy and D. Bhattacharjee, "A novel quaternary pattern of local maximum quotient for heterogeneous face recognition," *Pattern Recognit. Lett.*, vol. 113, pp. 19–28, Oct. 2018.

[31] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[32] 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.

[33] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2000–2006.

[34] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1679–1686.

[35] 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.

[36] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 687–694.

[37] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.

[38] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 1005–1010.

[39] X. Gao, J. Zhong, J. Li, and C. Tian, "Face sketch synthesis algorithm based on E-HMM and selective ensemble," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 4, pp. 487–496, Apr. 2008.

[40] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Sep. 2008.

[41] H. Zhou, Z. Kuang, and K. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1091–1097.

[42] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.

[43] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.

[44] N. Wang, J. Li, D. Tao, X. Li, and X. Gao, "Heterogeneous image transformation," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 77–84, Jan. 2013.

[45] N. Wang, X. Gao, and J. Li, "Random sampling for fast face sketch synthesis," *Pattern Recognit.*, vol. 76, pp. 215–227, Apr. 2018.

[46] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Sep. 2015.

[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[49] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 100–107.

[50] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "TV-GAN: Generative adversarial network based thermal to visible face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 174–181.

[51] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7355–7362.

[52] B. Cao, N. Wang, J. Li, and X. Gao, "Data augmentation-based joint learning for heterogeneous face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1731–1743, Jun. 2019.

[53] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1857–1865.

[54] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[55] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, Jun. 2005, pp. 539–546.

[56] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

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

[58] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[59] 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.

[60] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[61] 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.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[65] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[66] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[67] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

HAN BYEOL BAE received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, the B.S. degree in information and communication engineering from Yonsei University, Wonju, in 2010, and the M.S. degree in biometrics engineering from Yonsei University, in 2015, where he is currently pursuing the Ph.D. degree with the Image and Video Pattern Recognition Laboratory. His research interests are mainly in face recognition, image translation, and image classification.

TAEJAE JEON received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2014, where he is currently pursuing the Ph.D. degree with the Image and Video Pattern Recognition Laboratory. His research interests include video classification, facial landmark detection, and stress recognition using deep learning.

YONGJU LEE received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2018, where he is currently pursuing the Ph.D. degree with the Image and Video Pattern Recognition Laboratory. His research interests include face recognition and facial landmark detection.

SUNGJUN JANG received the B.S. degree in electronics engineering from Kwangwoon University, Seoul, South Korea, in 2019. He is currently pursuing the M.S. degree with the Image and Video Pattern Recognition Laboratory, Yonsei University. His research interests include face recognition and semantic segmentation using deep learning.

SANGYOUN LEE (Member, IEEE) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1999. He is currently a Professor and the Head of electrical and electronic engineering with the Graduate School and the Head of the Image and Video Pattern Recognition Laboratory, Yonsei University. His research interests include all aspects of computer vision, with a special focus on pattern recognition for face detection and recognition, advanced driver-assistance systems, and video codecs.