A Unified Framework for Bidirectional Prototype Learning from Contaminated Faces across Heterogeneous Domains

Meng Pang, Binghui Wang, *Member, IEEE,* Siyu Huang, Yiu-ming Cheung, *Fellow, IEEE,* and Bihan Wen*, *Member, IEEE*

***Abstract*—Existing heterogeneous face synthesis (HFS) methods focus on performing accurate image-to-image translation across domains, while they cannot effectively remove the nuisance facial variations such as poses, expressions or occlusions. To address such challenges, this paper studies a new practical heterogeneous prototype learning (HPL) problem. To be specific, given a face image contaminated by facial variations from a source domain, HPL aims to reconstruct the variation-free prototype in a specified target domain. To tackle HPL, we propose a unified and end-to-end framework named bidirectional heterogeneous prototype learning (BHPL). As a bidirectional learning framework, BHPL is able to simultaneously reconstruct the heterogeneous prototypes across *source-to-target* as well as *target-to-source* domains. Furthermore, BHPL is capable of learning the identity prototype features for the contaminated face images from both source and target domains in order to perform robust heterogeneous face recognition. BHPL consists of an encoder-decoder stuctural generator and two dual-task discriminators, which play an adversarial game such that the generator learns the identity prototype feature and generates the cross-domain identity-preserved prototype for each input face image from both domains, and the discriminators accurately predict face identity and distinguish real versus fake prototypes. Empirically studies on multiple heterogeneous face datasets containing facial variations demonstrate the effectiveness of BHPL.**

***Index Terms*—Face synthesis, heterogeneous prototype learning, heterogeneous face recognition, adversarial learning.**

I. Introduction

HETEROGENEOUS face synthesis (HFS), i.e., translating a face image from the source domain to the target domain via image synthesis, has attracted wide attentions in artificial intelligence security owing to its potential applications in criminal identification, law enforcement, access control, digital entertainment, person re-identification, to name a few [1–15]. To tackle HFS, a number of reconstruction-based methods [16–20] and deep generative model-based methods [21–25] have been developed. Generally, these methods assume that the input image in the source domain is *clean*; and aim at transferring its domain style, e.g., from near infrared (NIR) to visible spectrum (VIS), while keeping the facial details unaltered in the target domain.

However, in real-world scenarios, the captured source domain face images are probably contaminated by various facial variations such as expressions, poses, misalignments, and occlusions. In such cases, the existing HFS methods [15–20, 22–24, 26, 27] simply transfer the domain style of images but ignore removing the facial variations, thus making the personal identity of the synthesized image in the target domain hard to be recognized by forensic experts. Therefore, it is important to generate an appropriate *variation-free* prototype[1] across the source-to-target heterogeneous domains to better represent the

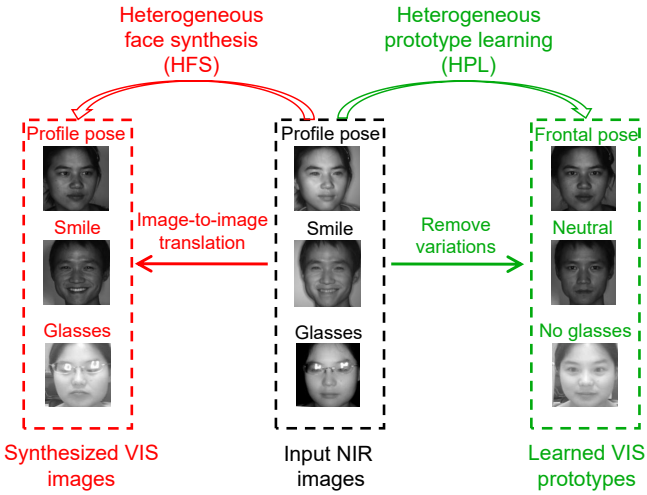


Fig. 1. Differences between the HPL and the classic HFS problems for NIR-to-VIS image synthesis. In HFS, the synthesized VIS images still contain the variations of pose, expressions, or occlusion that exist in the input NIR images. In HPL, the learned VIS prototypes are variation-free and preserve the personal identity characteristics in the input NIR images.

Meng Pang and Bihan Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798 Singapore e-mails: (meng.pang, bihan.wen)@ntu.edu.sg. Meng Pang and Bihan Wen were supported in part by Ministry of Education, Republic of Singapore under the start-up grant, the National Research Foundation (NRF) Singapore and the Singapore Cybersecurity Consortium (SGCSC) Grant Office, under grant SGCSC_Grant_2019-S01. Bihan Wen is the corresponding author.

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, e-mail: ymc@comp.hkbu.edu.hk. Yiu-ming Cheung was supported in part by the NSFC/RGC Joint Research Scheme under grant: N_HKBU214/21, RGC General Research Fund under grant: 12201321, the NSFC under grant: 61672444, Hong Kong Baptist University under grants: RC-FNRA-IG/18-19/SCI/03 and RC-IRCMs/18-19/SCI/01, and the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Project ITS/339/18.

Binghui Wang is with the Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois, USA, e-mail: bwang70@iit.edu. Binghui Wang was supported by his Startup funding.

Siyu Huang is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, MA 02134, USA, e-mail: huang@seas.harvard.edu.

[1]A prototype indicates a frontal face image with neutral expression, under normal illumination, and without occlusion.

personal identity. Such a new and practical issue is termed as *heterogeneous prototype learning (HPL)*. Different from the classic HFS that simply executes image-to-image translation, HPL aims to simultaneously remove the facial variations and preserve the personal identity during domain transferring. For a better clarification, we take the NIR to VIS image synthesis as an example and compare HPL with HFS in Fig. 1. It is clear that the existing HFS methods are unsuitable for HPL, as their synthesized VIS images would still contain variations such as pose, expression (e.g., smile), or occlusion (e.g., wearing glasses) that exist in the input NIR images although the domain style is transferred. We also find that most existing prototype learning-based methods [28–31] cannot be directly applied to HPL as they focus on generating the homogeneous prototypes in a *single* domain.

To address HPL, a straightforward idea is to perform domain transfer and prototype learning sequentially (or vice versa) in a two-stage process. However, we argue that such naive two-stage solution is far less satisfactory as the involved two sub-problems, i.e., prototype learning and domain transfer, are naturally cross-coupled in HPL. Particularly, when a synthesized face image or learned prototype has certain distortion, such distortion would be largely magnified when propagating to the other stage. Consequently, it motivates us to seek a *systematic* solution to HPL that is able to solve the two sub-problems jointly using a unified and end-to-end framework.

In order to learn the heterogeneous prototype across the source-to-target domains, we advocate treating the source domain style (e.g., texture) information as a special type of variation in the input image, while the target domain style information as a crucial component of the desired learned prototype. In this manner, we convert the union of the above-mentioned two sub-problems in HPL into a *generalized* prototype learning issue. Furthermore, given that the heterogeneous prototype learning is bidirectional by nature, i.e., the mappings between two domains are always coupled, we can simultaneously learn the prototype across the target-to-source domains, which could better preserve personal identity through the opposite learning direction and avoid re-training once the source and target domains are switched. In this paper, we therefore propose a unified and end-to-end framework based on adversarial learning, namely bidirectional heterogeneous prototype learning (BHPL). As illustrated in Fig. 2, BHPL consists of an encoder-decoder structural generator and two dual-task discriminators. The generator G is a symmetric network possessing two pathways, with each having an encoder-decoder subnet that 1) extracts the identity prototype feature and 2) generates the cross-domain prototype, from a source domain (or target domain) contaminated input image. The two discriminators D and $\tilde{D}$ compete with G to enforce: 1) the learned identity prototype features from both the source and target domain input images to encode as much identity information as possible, which can be applied to perform heterogeneous face recognition (HFR); and 2) the learned cross-domain prototype in each pathway is variation-free, which accurately captures the personal characteristics of the source domain (or target domain) input image.

To the best of our knowledge, the proposed BHPL is the first attempt to address the HPL problem using a unified and end-to-end framework. As a bidirectional learning framework, BHPL is able to simultaneously learn the heterogeneous prototypes across the source-to-target as well as the target-to-source domains. Furthermore, BHPL learns the identity prototype features for the contaminated face images from both source and target domains so as to perform HFR. The contributions of this work are summarized as follows:

- We propose the novel BHPL to address the practical heterogeneous prototype learning problem, i.e., reconstructing the variation-free prototype from a contaminated face image across heterogeneous domains, using a unified and end-to-end framework.
- We design a symmetric encoder-decoder structural generator, which simultaneously learns the heterogeneous prototypes across the source-to-target and the target-to-source domains. Furthermore, the generator is capable of learning the identity prototype features of contaminated face images from both domains.
- We design two adversarial discriminators to assist the generator in removing the facial variations and meanwhile preserving the identity information of the contaminated face images in the generated cross-domain prototypes.
- We conduct extensive experiments on multiple real-world NIR-VIS, ID-camera, and photograph-sketch heterogeneous face datasets, to demonstrate the powerful capability of BHPL for heterogeneous prototype learning, as well as the effectiveness of the learned identity prototype features for HFR.

The reminder of this paper is organized as follows. Section II provides an overview of the related works, and Section III briefly reviews the generative adversarial network (GAN). Section IV introduces the proposed BHPL in details. In Section V, qualitative and quantitative experiments are conducted on four real-world heterogeneous face datasets to evaluate the performance of BHPL. In Section VI, we discuss the generalization ability, universality, and limitations of BHPL. Finally, we give the conclusion and future works in Section VII.

II. Related Work

A. Heterogeneous Face Synthesis

HFS aims to generate cross-domain face images and compare them in the same domain. In practice, the domain style could be light spectrum (e.g., near infrared), artistic style (e.g., sketch), resolution, etc. Liu *et al*. [16] firstly studied this problem and utilized the local linear embedding (LLE) [32] to preserve the local reconstruction structure during face synthesis. Wang *et al*. [33] employed the Markov random field (MRF) to characterize the relationships between neighboring face patches to meet the smoothness requirement. Xu *et al.* [34] proposed a cross-spectral dictionary learning approach to reconstruct pseudo-face images between the NIR and VIS domains. Subsequently, a series of reconstruction-based methods [17–20, 35] were developed to synthesize face image in the target domain based on a pre-defined or learned source domain patch dictionary.

Recently, deep learning has seen rapid development and received increased attention in cross-domain image synthesis, image matching, and representation learning [36]. Zhang *et al*. [37] proposed a fully convolutional network (FCN) with a joint generative discriminative minimization objective to learn the photograph-sketch mapping. Lezama *et al.* [21] adopted a deep neural network to synthesize VIS images from NIR images followed by a low-rank embedding enhancement step. Zhang *et al.* [10] proposed a deep high-resolution pseudo-siamese framework for cross-resolution image matching and achieved state-of-the-art performance. Furthermore, motivated by the success of GAN [38] in image synthesis and style transferring, a number of deep generative model-based techniques [2, 22–24, 39] have been proposed to tackle HFS. Isola *et al*. [23] released a Pix2Pix software based on conditional GAN for image translation. Zhu *et al*. [22] proposed a cycle consistent GAN (cycle-GAN) to learn heterogeneous face image across different domains using unpaired training data. Zhao *et al.* [27] proposed an adversarial consistency loss-based GAN (ACL-GAN) for image translation to encourage the translated images to retain important features of the source images. Liu *et al*. [40] proposed an unsupervised image-to-image translation (UNIT) framework provided that a pair of corresponding images from different domains can be mapped to a same representation in a shared-latent space. Fu *et al*. [2] proposed a dual generation model to generate massive paired NIR-VIS images from noise and thus reducing the domain gap for heterogeneous face recognition. Zhang *et al.* [39] proposed a multidomain adversarial learning (MDAL) for photograph-sketch synthesis by learning the reconstruction process in each domain. Song *et al*. [24] incorporated feature learning into HFS and proposed an adversarial discriminative feature learning (ADFL) approach that performs adversarial learning on both spatial and feature spaces.

Lately, Yu *et al.* [41] developed a pose-preserving cross-spectral face hallucination framework consisting of an attention warping module and a mutual information constraint, to alleviate the misalignment in synthesized images. To better preserve the texture information, Duan *et al*. [25] proposed a pose agnostic cross-spectral hallucination (PACH) framework by introducing a texture prior synthesis module. Although the two methods can deal with the misalignments or certain pose variations during image synthesis, they cannot generalize many other facial variations such as expressions or occlusions.

B. Prototype Learning

Prototype learning is an emerging hot topic which aims to reconstruct the prototype (i.e., standard face image) for a contaminated enrolment sample with facial variations such as expressions, poses and occlusions. In the literature, there are two types of prototype learning-based methods [31]: one is to exploit auxiliary information from query set for image restoration, the other is to train appropriate mappings between contaminated and standard samples.

For the former type, Gao *et al*. [42] and Pang *et al*. [28] proposed to estimate the prototypes by the clustering centroid of the union of the labeled enrolment and unlabeled query sets via Gaussian mixture model (GMM) [43] or a semi-supervised low-rank representation. Despite promising prototypes obtained by them, they need to acquire the unknown query set in advance, which may be impractical from a real-time perspective. For the latter type, benefiting from the powerful mapping ability of GAN, a series of GAN variants [29–31, 44] have been developed to decrease facial variations in contaminated samples and to synthesize the corresponding identity-preserved prototypes. Song *et al*. [29] proposed a geometry-guided GAN by using fiducial points to guide facial expression transfer and neutralization. Chen *et al*. [30] proposed an occlusion-aware GAN to detect and recover missing regions in occluded face samples. Huang *et al*. [44] proposed a two-pathway GAN to correct the ill-posed samples through both global and local transformations. More recently, Pang *et al*. [31] proposed a general variation disentangling GAN framework to handle universal facial variations. However, these GAN variants cannot be applied to HPL because they ignore the domain differences during prototype generation. By contrast, our proposed BHPL treats the source domain style information as a special type of variation while the target domain style information as a crucial component of the desired learned prototype. In doing so, BHPL converts HPL into a generalized prototype learning issue for solving.

III. Review on GAN

Goodfellow *et al*. [38] proposed GAN to train a generative model for image synthesis. It consists of two key components, i.e., a generator G and a discriminator D, which compete in a two-player minimax game. The discriminator D is trained to distinguish between the real image $\mathbf{x}$ and the fake generated image $\hat{\mathbf{x}}$, while the generator G is trained to generate realistic-looking image, i.e., $G(\mathbf{z})$, based on a random noise vector $\mathbf{z}$ to fool D. $\mathbf{x}$ and $\mathbf{z}$ are sampled from their respected distributions p_{data} and p_z, i.e., $\mathbf{x} \sim p_{data}$, $\mathbf{z} \sim p_z$. Concretely, the objective function of GAN is presented as follows:

$$\min_G \max_D V = E_{\mathbf{x}}[\log D(\mathbf{x})] + E_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]. \quad (1)$$

It has been proved that this minimax game in Eq. (1) has a global optimum when the distribution of the generated images approaches to the distribution of the real images [38]. Furthermore, to provide stronger gradients early in learning, Goodfellow *et al.* suggested to replace the minimization of $\log(1 - D(G(\mathbf{z})))$ with the maximization of $\log(D(G(\mathbf{z})))$. Therefore, Eq. (1) can be reformulated as follows:

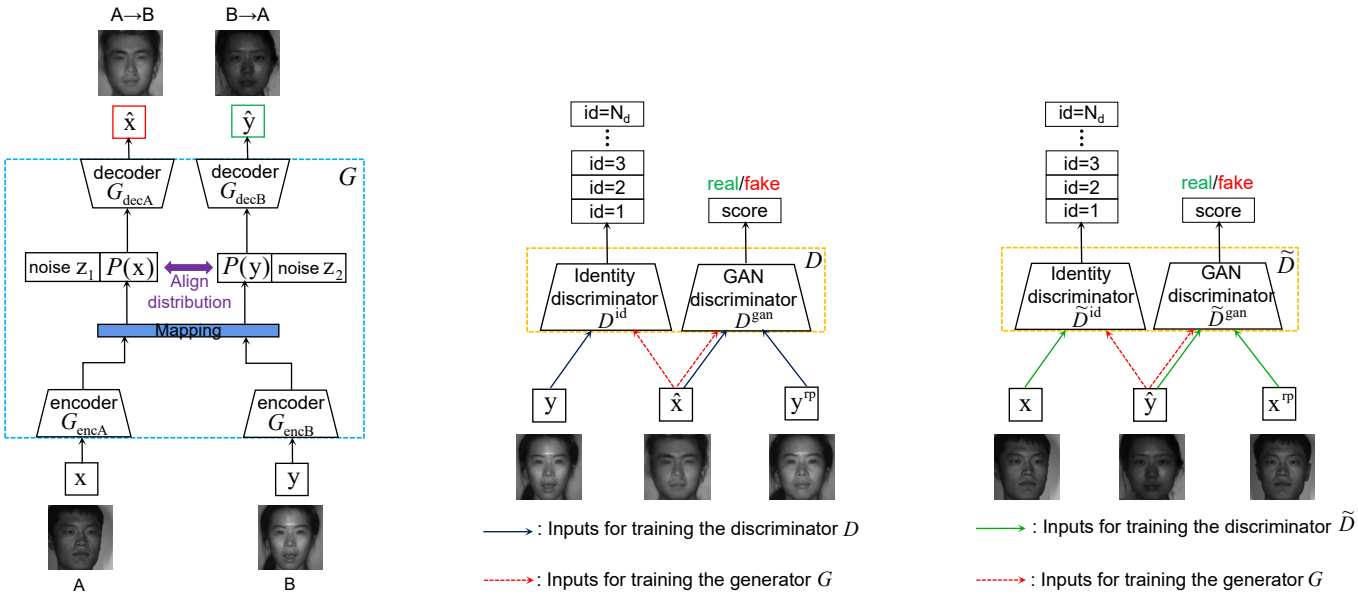
$$\max_D V_D = E_{\mathbf{x}}[\log D(\mathbf{x})] + E_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))], \quad (2)$$

$$\max_G V_G = E_{\mathbf{z}}[\log(D(G(\mathbf{z})))]. \quad (3)$$

The discriminator D in Eq. (2) and the generator G in Eq. (3) are iteratively updated until convergence is achieved or a predefined maximum number iterations is reached.

IV. The Proposed Framework

In this section, we first define the problem and the objectives. Then, we introduce the architecture of the proposed BHPL, followed by the training scheme and applications. For clarity, we summarize some important notations in Table I.



(a) Generator G (b) Discriminator D (c) Discriminator $\tilde{D}$

Fig. 2. Overview of the proposed BHPL. $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}^{rp}$, $\mathbf{y}^{rp}$, $\widehat{\mathbf{x}}$, and $\widehat{\mathbf{y}}$ denote the input image from Domain A, the input image from Domain B, the real Domain A prototype, the real Domain B prototype, the learned Domain B prototype of $\mathbf{x}$, and the learned Domain A prototype of $\mathbf{y}$, respectively. $P(\mathbf{x})$ and $P(\mathbf{y})$ denote the learned identity prototype features of $\mathbf{x}$ and $\mathbf{y}$, respectively. When training D, D^{id} predicts the identity label of $\mathbf{y}$; D^{gan} assigns a high score to $\mathbf{y}^{rp}$, and low score to $\widehat{\mathbf{x}}$. When training $\tilde{D}$, $\tilde{D}^{id}$ predicts the identity label of $\mathbf{x}$; $\tilde{D}^{gan}$ assigns a high score to $\mathbf{x}^{rp}$, and low score to $\widehat{\mathbf{y}}$. When training G, $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$ enable D^{id} and $\tilde{D}^{id}$ to classify them into the identity labels of $\mathbf{x}$ and $\mathbf{y}$, respectively, and fool D^{gan} and $\tilde{D}^{gan}$ to assign them high scores; in addition, the feature distributions of $P(\mathbf{x})$ and $P(\mathbf{y})$ are aligned in order to reduce their domain discrepancy.

A. Problem Definition

We propose to jointly perform heterogeneous prototype learning and feature learning for contaminated face images across domains using a unified and end-to-end framework.

To be specific, suppose a training set consists of N_d identities from both Domain A and Domain B. The training set is allowed to be unpaired, i.e., the samples in Domain A and in Domain B are not one-to-one. Each image $\mathbf{x}$ in Domain A is labeled by $l_x = \{l_x^{id}, l_x^{var}\}$; while in Domain B, each image $\mathbf{y}$ is labeled by $l_y = \{l_y^{id}, l_y^{var}\}$. l_x^{id} (or l_y^{id}) represents the face identity of $\mathbf{x}$ (or $\mathbf{y}$). l_x^{var} (or l_y^{var}) indicates whether $\mathbf{x}$ (or $\mathbf{y}$) contains arbitrary facial variations. Taking $\mathbf{x}$ for example, if $\mathbf{x}$ has a variation (e.g., expression, pose, illumination, or occlusion), then $l_x^{var} = 1$; otherwise $l_x^{var} = 0$. We denote that each $\mathbf{x}$ in Domain A is sampled from the distribution $\mathcal{P}_{dataA}$, i.e., $\mathbf{x} \sim \mathcal{P}_{dataA}$, and each $\mathbf{y}$ in Domain B from the distribution $\mathcal{P}_{dataB}$, i.e., $\mathbf{y} \sim \mathcal{P}_{dataB}$. Given a testing contaminated image in Domain A, denoted as $\mathbf{x}_t$, and a testing contaminated image in Domain B, denoted as $\mathbf{y}_t$, BHPL aims to achieve the following two crucial objectives:

- **Heterogeneous prototype learning**: Learning a proper Domain B prototype $\widehat{\mathbf{x}}_t$ for $\mathbf{x}_t$ and Domain A prototype $\widehat{\mathbf{y}}_t$ for $\mathbf{y}_t$, such that $\widehat{\mathbf{x}}_t$ (or $\widehat{\mathbf{y}}_t$): 1) is variation-free, and 2) preserves the identity characteristics of $\mathbf{x}_t$ (or $\mathbf{y}_t$).
- **Identity prototype feature learning**: Learning a discriminative identity prototype feature $P(\mathbf{x}_t)$ for $\mathbf{x}_t$ and $P(\mathbf{y}_t)$ for $\mathbf{y}_t$, such that $P(\mathbf{x}_t)$ (or $P(\mathbf{y}_t)$) represents the identity of $\mathbf{x}_t$ (or $\mathbf{y}_t$) accurately.

B. BHPL Architecture

We achieve the above two objectives by proposing a joint heterogeneous prototype learning and feature learning framework, i.e., BHPL, whose architecture is presented in Fig. 2. BHPL consists of three main modules: an encoder-decoder structural generator G and two dual-task discriminators D and $\tilde{D}$. D and $\tilde{D}$ compete with G that enforce: 1) the learned identity prototype features from both the source and target domain input images to encode as much identity information as possible; and 2) the learned cross-domain prototype in each pathway is variation-free and captures the personal identity characteristics accurately. In the following, we will introduce the generator and discriminators in details.

1) Generator G: The generator G is composed of two encoders, i.e., G_{encA} and G_{encB}, and two decoders, i.e., G_{decA} and G_{decB}. G_{encA} encodes an identity prototype feature $P(\mathbf{x})$ for an input Domain A image $\mathbf{x}$; while G_{encB} encodes an identity prototype feature $P(\mathbf{y})$ for an input Domain B image $\mathbf{y}$. G_{decA} and G_{decB} take the concatenation of $P(\mathbf{x})$ with the noise $\mathbf{z}_1$ and the concatenation of $P(\mathbf{y})$ with the noise $\mathbf{z}_2$ as the inputs, and then generate a Domain B prototype $\widehat{\mathbf{x}} = G_{decA}(P(\mathbf{x}), \mathbf{z}_1)$ for $\mathbf{x}$, and a Domain A prototype $\widehat{\mathbf{y}} = G_{decB}(P(\mathbf{y}), \mathbf{z}_2)$ for $\mathbf{y}$, respectively. The noises $\mathbf{z}_1$ and $\mathbf{z}_2$ are draw from uniform distribution $[-1, 1]^{Nz}$.

2) Discriminators D and $\tilde{D}$: The discriminator D is a dual-task discriminator involving two sub-discriminators, namely D^{id} and D^{gan}. To be specific,

1) D^{id} is an identity-related sub-discriminator that outputs a vector of N_d-dimension for face identity classification. N_d denotes the total number of training identities.
2) D^{gan} is a GAN-related sub-discriminator that distinguishes the real versus fake prototypes in Domain B. It assigns a score to the image and a higher score indicates that the image is closer to the real prototype.

TABLE I
MEANING OF THE NOTATIONS IN BHPL.

Notation	Meaning
$\mathbf{x}$	Image in Domain A, $\mathbf{x} \sim \mathcal{P}_{dataA}$
$\mathbf{y}$	Image in Domain B, $\mathbf{y} \sim \mathcal{P}_{dataB}$
$\mathbf{x}^{rp}$	Real prototype in Domain A, $\mathbf{x}^{rp} \sim \mathcal{P}_{realA}$
$\mathbf{y}^{rp}$	Real prototype in Domain B, $\mathbf{y}^{rp} \sim \mathcal{P}_{realB}$
l_x^{id}/l_x^{var}	The identity/variation label for $\mathbf{x}$
l_y^{id}/l_y^{var}	The identity/variation label for $\mathbf{y}$
G	The encoder-decoder structural generator
G_{encA}	The encoder A in G for $\mathbf{x}$
G_{encB}	The encoder B in G for $\mathbf{y}$
$P(\mathbf{x})$	The encoded identity prototype feature of $\mathbf{x}$
$P(\mathbf{y})$	The encoded identity prototype feature of $\mathbf{y}$
G_{decA}	The decoder A in G
G_{decB}	The decoder B in G
$\widehat{\mathbf{x}}$	The generated Domain B prototype of $\mathbf{x}$
$\widehat{\mathbf{y}}$	The generated Domain A prototype of $\mathbf{y}$
$D, \tilde{D}$	The multi-task discriminators
D^{gan}, $\tilde{D}^{gan}$	The discriminators to distinguish prototypes
D^{id}, $\tilde{D}^{id}$	The discriminators to classify face identity

Similar to D, $\tilde{D}$ is also a dual-task discriminator consisting of two sub-discriminators $\tilde{D}^{id}$ and $\tilde{D}^{gan}$. $\tilde{D}^{id}$ still outputs a N_d-dimensional vector for classifying identity labels, while $\tilde{D}^{gan}$ is designed to distinguish real versus fake prototypes in Domain A.

C. BHPL Training

1) Training of G: For the generator G, it has the following three objectives:

- Fool D^{gan} to classify $\widehat{\mathbf{x}}$ as real Domain B prototype, and $\tilde{D}^{gan}$ to classify $\widehat{\mathbf{y}}$ as real Domain A prototype.
- Enable D^{id} to classify $\widehat{\mathbf{x}}$ as the same identity label as $\mathbf{x}$ (i.e., l_x^{id}), and $\tilde{D}^{id}$ to classify $\widehat{\mathbf{y}}$ as the same identity label as $\mathbf{y}$ (i.e., l_y^{id}).
- Align the feature distributions of $P(\mathbf{x})$ and $P(\mathbf{y})$ to reduce the domain discrepancy.

By considering all the above objectives, the final objective function V_G for training G is presented as follows:

$$\max_G V_G = V_G^{gan} + \alpha_1 V_G^{id} - \alpha_2 V_G^{dis}, \tag{4}$$

where α_1 and α_2 are two positive trade-off parameters. V_G^{gan}, V_G^{id}, and V_G^{dis} are defined as follows:

$$V_G^{gan} = E_{\mathbf{x},\mathbf{y}}[\log D^{gan}(\widehat{\mathbf{x}}) + \log \tilde{D}^{gan}(\widehat{\mathbf{y}})], \tag{5}$$

$$V_G^{id} = E_{\mathbf{x},\mathbf{y}}[\log D^{id}_{l_x^{id}}(\widehat{\mathbf{x}}) + \log \tilde{D}^{id}_{l_y^{id}}(\widehat{\mathbf{y}})], \tag{6}$$

$$V_G^{dis} = \mathrm{MMD}^2(\mathcal{P}_{feaA}, \mathcal{Q}_{feaB}), \tag{7}$$

where D_i^{id} is the i-th element in D^{id}, $\tilde{D}_j^{id}$ is the j-th element in $\tilde{D}^{id}$, $\mathcal{P}_{feaA}$ and $\mathcal{Q}_{feaA}$ denote the distributions of $P(\mathbf{x})$ and $P(\mathbf{y})$, respectively. In Eq. (7), we minimize the squared maximum mean discrepancy (MMD) distance to reduce the divergence between $\mathcal{P}_{feaA}$ and $\mathcal{Q}_{feaB}$. This aims to generate two feature distributions in the latent space that are identical, and thus force to exclude the domain information. Theoretically, MMD reaches its global minimum zero if and only if the two distributions are equal. For more details about the MMD metric, please refer to [45].

2) Training of D and $\tilde{D}$: Subsequently, according to the values of l_x^{var} and l_y^{var}, we collect standard uncontaminated face images from Domain A and Domain B in the training set to construct the real Domain A and real Domain B prototype corpuses, respectively. We denote each real Domain A prototype as $\mathbf{x}^{rp} \sim \mathcal{P}_{realA}$, and real Domain B prototype as $\mathbf{y}^{rp} \sim \mathcal{P}_{realB}$.

For the discriminator $D = [D^{gan}, D^{id}]$, it has the following two objectives:

- Given the *real* Domain B prototype $\mathbf{y}^{rp}$ and the generated *fake* Domain B prototype $\widehat{\mathbf{x}}$, D^{gan} aims to classify $\mathbf{y}^{rp}$ as the real prototype and classify $\widehat{\mathbf{x}}$ as the fake one.
- Given the input Domain B image $\mathbf{y}$, D^{id} aims to correctly predict its identity label l_y^{id}.

Formally, the final objective function V_D for training D is

$$\max_D V_D = V_D^{gan} + \beta V_D^{id}, \tag{8}$$

where β is a positive trade-off parameter, and the sub-objective functions V_D^{gan} and V_D^{id} are defined as

$$V_D^{gan} = E_{\mathbf{y}^{rp}}[\log D^{gan}(\mathbf{y}^{rp})] + E_{\mathbf{x}}[\log(1 - D^{gan}(\widehat{\mathbf{x}}))], \tag{9}$$

$$V_D^{id} = E_{\mathbf{y}}[\log D^{id}_{l_y^{id}}(\mathbf{y})]. \tag{10}$$

For the other discriminator $\tilde{D} = [\tilde{D}^{gan}, \tilde{D}^{id}]$, it also has two objectives:

- Given the *real* Domain A prototype $\mathbf{x}^{rp}$ and the generated *fake* Domain A prototype $\widehat{\mathbf{y}}$, $\tilde{D}^{gan}$ aims to classify $\mathbf{x}^{rp}$ as the real prototype and classify $\widehat{\mathbf{y}}$ as the fake one.
- Given the input Domain A image $\mathbf{x}$, $\tilde{D}^{id}$ aims to correctly predict its identity label l_x^{id}.

Formally, the final objective function $V_{\tilde{D}}$ for training $\tilde{D}$ is

$$\max_{\tilde{D}} V_{\tilde{D}} = V_{\tilde{D}}^{gan} + \gamma V_{\tilde{D}}^{id}, \tag{11}$$

where γ is a positive trade-off parameter, and $V_{\tilde{D}}^{gan}$ and $V_{\tilde{D}}^{id}$ are defined as

$$V_{\tilde{D}}^{gan} = E_{\mathbf{x}^{rp}}[\log \tilde{D}^{gan}(\mathbf{x}^{rp})] + E_{\mathbf{y}}[\log(1 - \tilde{D}^{gan}(\widehat{\mathbf{y}}))], \tag{12}$$

$$V_{\tilde{D}}^{id} = E_{\mathbf{x}}[\log \tilde{D}^{id}_{l_x^{id}}(\mathbf{x})]. \tag{13}$$

The training procedure of BHPL is presented in **Algorithm 1**. It can be seen that, we alternatively train the generator G, the discriminator D, and the discriminator $\tilde{D}$ by solving the objective functions V_G in Eq. (4), V_D in Eq. (8), and $V_{\tilde{D}}$ in Eq. (11) iteratively. During the alternative training process, G, $D = [D^{gan}, D^{id}]$, and $\tilde{D} = [\tilde{D}^{gan}, \tilde{D}^{id}]$ will be updated and improved. To be specific,

- With D^{gan} and $\tilde{D}^{gan}$ being more powerful in distinguishing real versus fake prototypes, G strives for generating the realistic-looking Domain B prototype $\widehat{\mathbf{x}}$ to fool D^{gan}, and Domain A prototype $\widehat{\mathbf{y}}$ to fool $\tilde{D}^{gan}$.
- Besides, D^{id} enables $\widehat{\mathbf{x}}$ to preserve the identity characteristics of $\mathbf{x}$, and $\tilde{D}^{id}$ enables $\widehat{\mathbf{y}}$ to preserve the identity

Algorithm 1 BHPL Training

Input: A training set of N_d identities from Domain A and Domain B with each image $\mathbf{x}$ (or $\mathbf{y}$) annotated by $l_x = \{l_x^{id}, l_x^{var}\}$ (or $l_y = \{l_y^{id}, l_y^{var}\}$); A real prototype corpus in Domain A with each image $\mathbf{x}^{rp}$ sampled from the distribution $\mathcal{P}_{realA}$; A real prototype corpus in Domain B with each image $\mathbf{y}^{rp}$ sampled from the distribution $\mathcal{P}_{realB}$.

1: **repeat**
2: Fix G, update G by solving the objective in Eq. (4)
3: Fix D, update D by solving the objective in Eq. (8)
4: Fix $\tilde{D}$, update $\tilde{D}$ by solving the objective in Eq. (11)
5: **until** convergence is achieved or a predefined maximum number of iterations is reached

Output: Trained G, D, $\tilde{D}$

characteristics of $\mathbf{y}$. Furthermore, D^{id} and $\tilde{D}^{id}$ guide G_{encA} and G_{encB} to learn the discriminative identity prototype features $P(\mathbf{x})$ and $P(\mathbf{y})$ that could encode as much identity information as possible.

- Moreover, minimizing MMD between $\mathcal{P}_{feaA}$ and $\mathcal{Q}_{feaB}$ further reduces the domain discrepancy.

It is worth mentioning that, BHPL is a bidirectional learning framework, which could avoid the re-training of the framework once the source and target domains are switched.

D. BHPL Applications

In testing, we can leverage our trained generator G to generate heterogeneous prototypes across domains as well as extract discriminative identity prototype features. Specifically, we can do the following two tasks:

1) **Heterogeneous prototype learning:** Generate an appropriate Domain B (or Domain A) prototype for a contaminated Domain A (or Domain B) face image to be recognized by forensic experts.
2) **Heterogeneous face recognition:** Given a Domain B (or Domain A) query face image and Domain A (or Domain B) enrolment set, we can obtain their discriminative identity prototype features and then perform classification. For simplicity, in the following experiments, we adopt a cosine distance-based nearest neighbor classifier for classification [2].

We will demonstrate the effectiveness of BHPL regarding the above applications in the subsequent experimental section.

V. Experimental Results

In this section, we first introduce the evaluated datasets, the implementation details of BHPL, and the parameter settings in Subsection V-A. Subsequently, we qualitatively and quantitatively evaluate our proposed BHPL by conducting the following experiments:

1) In Subsections V-B-V-D, we evaluate the learned heterogeneous prototypes and features by BHPL on two NIR-VIS face datasets, i.e., BUAA NIR-VIS and CASIA NIR-VIS v2.0, on one ID-camera face dataset, i.e., NJU-ID, and on one photograph-sketch face dataset, i.e., CUFSF, respectively.
2) In Subsection V-E, we make an comparison between the learned heterogeneous prototypes by BHPL and the synthesized heterogeneous images by the existing advanced HFS approaches on the above datasets.
3) In Subsection V-F, we perform ablation study to investigate the roles of the identity-related sub-discriminator, GAN-related sub-discriminator, and the the MMD constraint on the performance of BHPL.

[2]There are some other distance metrics such as l_1-distance and l_2-distance that can be used. Through experiments, we observe their performance are comparable or inferior to that of cosine distance.

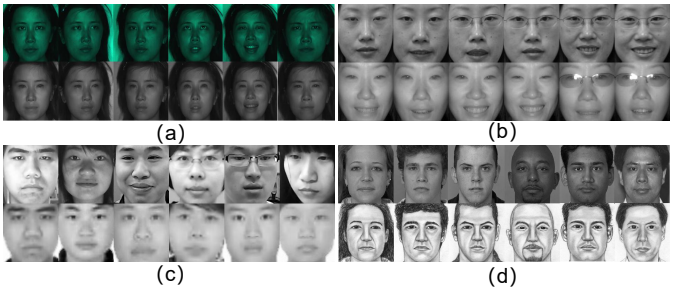


Fig. 3. Example images from four heterogeneous face datasets: (a) BUAA NIR-VIS; (b) CASIA NIR-VIS v2.0; (c) NJU-ID; (d) CUFSF.

A. Experimental Setting

1) Dataset Description: BUAA NIR-VIS [46] consists of 150 identities with each having 9 VIS and 9 NIR face images. In the experiments, a total of 50 identities with 900 images are randomly chosen as the training set, while the remaining 100 identities with 1800 images as the testing set.

CASIA NIR-VIS v2.0 [47] is the second edition of the CASIA NIR-VIS dataset. It contains 725 identities with each having 5-50 NIR face images and 1-22 VIS face images. We perform experiments using the standard protocol as the existing methods did. For each fold, we select 360 identities containing about 2500 VIS and 6100 NIR images for training and choose another 358 identities as the testing set.

NJU-ID [48] is built for studying the ID-camera face recognition/verification. NJU-ID consists of 256 identities, with each having one low-resolution ID card image and one high-resolution photo image collected from laptop camera. We randomly select 100 identities for training, while the rest 156 ones are used for testing.

CUFSF [49] is a widely-used viewed sketch dataset for photograph to sketch synthesis and recognition. It consists of 1194 identities from FERET dataset [50] with each having one standard photograph and one sketch drawn by an artist. In the experiments, we adopt two settings to evaluate the learned heterogeneous prototype and learned identity prototype feature, respectively. Firstly, we randomly select 200 identities and expand the photograph size by introducing samples from five subsets on FERET (bj, bk, bd, bf and bg) including variations of poses and expressions, thus each identity having 6 photographs and 1 sketch image. We use the first 100 identities for training while the rest 100 ones for testing. Secondly, we follow the protocol in [20] to randomly select 550 identities

TABLE II
THE NETWORK STRUCTURES OF D (OR $\tilde{D}$) AND G_{decA} (OR G_{decB}). F AND S DENOTE THE FILTER AND STRIDE, RESPECTIVELY.

D or $\tilde{D}$			G_{decA} or G_{decB}		
Layer	F/S	Output Size	Layer	F/S	Output Size
			FC	–	8×8×256
Conv1	3×3/1	128×128×32	DeConv1	3×3/1	8×8×160
Conv2	3×3/1	128×128×64	DeConv2	3×3/1	8×8×256
Conv3	3×3/2	64×64×64	DeConv3	3×3/2	16×16×256
Conv4	3×3/1	64×64×64	DeConv4	3×3/1	16×16×128
Conv5	3×3/1	64×64×128	DeConv5	3×3/1	16×16×192
Conv6	3×3/2	32×32×128	DeConv6	3×3/1	32×32×192
Conv7	3×3/1	32×32×96	DeConv7	3×3/1	32×32×96
Conv8	3×3/1	32×32×192	DeConv8	3×3/1	32×32×128
Conv9	3×3/2	16×16×192	DeConv9	3×3/2	64×64×128
Conv10	3×3/1	16×16×128	DeConv10	3×3/1	64×64×64
Conv11	3×3/1	16×16×256	DeConv11	3×3/1	64×64×64
Conv12	3×3/2	8×8×256	DeConv12	3×3/2	128×128×64
Conv13	3×3/1	8×8×160	DeConv13	3×3/1	128×128×32
Conv14	3×3/1	8×8×320	DeConv14	3×3/1	128×128×3
Pool	8×8/1	1×1×320			
FC	–	N_d+1			

TABLE III
DATASET PARTITION AND PARAMETER SETTING.

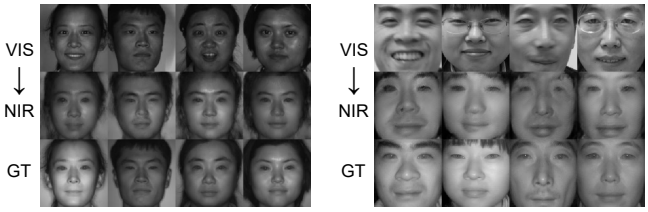
Dataset	#Training identity	#Testing identity	N_d	Trade-off parameter
BUAA NIR-VIS	50	100	50	α_1=β=γ=2 α_2=0.1
CASIA NIR-VIS v2.0	360	358	360	
NJU-ID	100	156	100	
CUFSF (setting 1)	100	100	100	
CUFSF (setting 2)	550	644	550	

containing 550 photograph-sketch pairs for training while the rest 644 identities for testing.

For each evaluated dataset, all face images are cropped to 128×128 pixels. Fig. 3 illustrates some face examples on BUAA NIR-VIS, CASIA NIR-VIS v2.0, NJU-ID, and CUFSF four heterogeneous datasets.

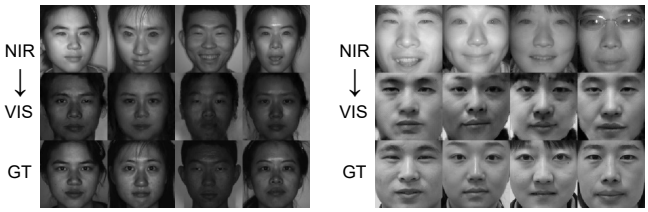
2) Implementation Details: For the decoders G_{decA} and G_{decB}, we adopt the CASIA-Net [51] as the backbone, where batch normalization (BN) and exponential linear unit (ELU) are used after each conv and deconv layer. D and $\tilde{D}$ both have an extra fully connection (FC) layer based on CASIA-Net, whose output is a (N_d+1)-dimensional vector for predicting the face identity and for distinguishing real versus fake prototypes. For clarity, the network structures of D (or $\tilde{D}$) and G_{decA} (or G_{decB}) are presented in Table II.

For the encoders G_{encA} and G_{encB}, we employ the Lightended CNN [52] pretrained on MS-Celeb-1M [53] as the backbone for feature extraction. Furthermore, for each evaluated dataset, we fine tune the Lightened CNN based on the corresponding training set. G extracts two 256-dimensional features via the two encoders and put them through a FC mapping layer, to generate the prototype feature $P(\mathbf{x}) \in \Re^{256}$ for $\mathbf{x}$ and $P(\mathbf{y}) \in \Re^{256}$ for $\mathbf{y}$. Subsequently, $P(\mathbf{x})$ is concatenated with the random noise $\mathbf{z}_1 \in \Re^{50}$, and fed to G_{decA} to generate the Domain B prototype of $\mathbf{x}$, i.e., $\widehat{\mathbf{x}}$; $P(\mathbf{y})$ is concatenated with the random noise $\mathbf{z}_2 \in \Re^{50}$, and fed to G_{decB} to generate the Domain A prototype of $\mathbf{y}$, i.e., $\widehat{\mathbf{y}}$.



(a) BUAA NIR-VIS (b) CASIA NIR-VIS v2.0

Fig. 4. Learned heterogeneous NIR prototypes by BHPL from four random VIS samples on (a) BUAA NIR-VIS and (b) CASIA NIR-VIS v2.0, respectively. Figures from top to bottom are: the input VIS samples, the learned NIR prototypes, and the groundtruth NIR prototypes.



(a) BUAA NIR-VIS (b) CASIA NIR-VIS v2.0

Fig. 5. Learned heterogeneous VIS prototypes by BHPL from four random NIR samples on (a) BUAA NIR-VIS and (b) CASIA NIR-VIS v2.0, respectively. Figures from top to bottom are: the input NIR samples, the learned VIS prototypes, and the groundtruth VIS prototypes.

We optimize the proposed BHPL model [3] by the mini-batch stochastic gradient descent (SGD) with a mini-batch size of 5. The maximum number of training epoches is set at 500. All weights are initialized from a zero-centered Normal distribution with the standard deviation of 0.02. Following the work in [51], we adopt the Adam optimizer [54] with tuned hyperparameters for optimizing, where the learning rate and momentum are set at 0.0002 and 0.5, respectively.

3) Parameter Setting: For the parameter setting, we denote N_d as the total number of identities in the training set, and N_z as the dimension of the random noise $\mathbf{z}_1$ and $\mathbf{z}_2$. The four trade-off hyper-parameters, i.e., α_1 and α_2 in Eq. (4), β in Eq. (8), and γ in Eq. (11), are tuned via grid search. Empirically, we observe that BHPL achieves promising performance when α_1, α_2, β, and γ are set at 2, 0.1, 2, and 2, respectively, and fix their values across all evaluated datasets. For clarity, we summarize the parameter settings and the training/testing sets partition on each dataset in Table III.

B. Evaluation on NIR-VIS Datasets

In this subsection, we evaluate BHPL on two challenging NIR-VIS face datasets, i.e., BUAA NIR-VIS and CASIA NIR-VIS v2.0. In the case, we treat VIS domain as Domain A, and NIR domain as Domain B.

Evaluation on learned prototype. With the trained BHPL model, we can obtain 1) the heterogeneous NIR prototypes for the VIS samples and 2) the heterogeneous VIS prototypes for the NIR samples, respectively. Then, we qualitatively measure the two types of heterogeneous prototypes. Fig. 4 and Fig. 5 show the learned heterogeneous NIR prototypes from four random VIS testing samples, and the learned heterogeneous

[3]The code is released at https://github.com/PangMeng92/BHPL_Codes.git.

TABLE IV
RECOGNITION ACCURACIES (%) OF BHPL AND THE OTHER COMPARED METHODS ON BUAA NIR-VIS AND CASIA NIR-VIS V2.0 HETEROGENEOUS DATASETS.

Methods		BUAA NIR-VIS	CASIA NIR-VIS v2.0
Hand-crafted	KDSR [55]	83.0	37.5
	CDFL [56]	–	71.5
	H2-LBP3 [57]	88.8	43.8
	CEFD [58]	–	85.6
	KMCM2L [48]	–	76.0
Deep learning	TRIVET [59]	93.9	95.7
	IDR [60]	94.3	97.3
	ADFL [24]	95.2	98.2
	DSU-Nets [61]	–	96.3
	PACH [25]	98.6	**98.9**
	RGM [62]	97.6	97.2
Ours	BHPL	**98.8**	97.3

TABLE V
RECOGNITION ACCURACIES (%) OF BHPL AND THE OTHER COMPARED METHODS ON NJU-ID DATASET.

Methods		Accuracy (%)
Face synthesis-based	MWF [18]	42.8
	Pix2Pix [23]	29.7
	RSLCR [20]	43.7
Metric learning-based	AJL-HFR [63]	86.4
	DA-JL [64]	89.0
Feature learning-based	CAL-HFR [65]	89.4
Ours	BHPL	**98.1**

VIS prototypes from four random NIR testing samples on each dataset, respectively. For reference, we also present the groundtruth (GT) NIR or VIS prototypes. From Fig. 4 and Fig. 5, we have the following key observations:

1) BHPL successfully reconstructs the *variation-free* NIR prototypes for the input VIS samples, as well as the VIS prototypes for the input NIR samples. Intuitively, for these VIS (or NIR) samples contaminated with facial variations of expressions (e.g., happiness and surprise), different poses, or occlusion of glasses, BHPL transfers the image style from VIS to NIR (or NIR to VIS) domain and meanwhile decreasing the corresponding variations.
2) For these learned heterogeneous NIR or VIS prototypes by BHPL, they preserve the personal identity characteristics well and look close to the groundtruth prototypes from the visual point.

Evaluation on learned feature. Following the established setting in the existing HFR methods, we choose one VIS sample from each identity in the testing set to form the enrolment set, and use all testing NIR samples in the testing set for querying. With the trained BHPL model, we can obtain the identity prototype features for both VIS enrolment and NIR query samples. Then, we quantitatively evaluate the learned features by applying them to perform HFR.

In the experiment, we choose 11 NIR-VIS feature learning-based methods for comparison, including 5 handcrafted feature learning-based HFR methods, i.e., kernelized discriminative spectral regression (KDSR) [55], coupled discriminative feature learning (CDFL) [56], H2-LBP3 [57], common encoding feature discriminant (CEFD) [58], and kernelized margin-based cross-modality metric learning (KMCM2L) [48], and 6 deep learning-based HFR methods, i.e., transfer NIR-VIS heterogeneous face recognition network (TRIVET) [59], invariant deep representation (IDR) [60], ADFL [24], domain specific units nets (DSU-Nets) [61], PACH [25], and relational deep feature learning (RGM) [62]. The rank-1 average recognition accuracies of all the methods on BUAA NIR-VIS and CASIA NIR-VIS v2.0 datasets are listed in Table IV. We have the following two key observations:

1) Compared to the traditional handcrafted feature learning based HFR methods, the deep learning-based HFR methods usually achieve better recognition performance. This indicates the good representation learning capability of deep neural networks.
2) Although our proposed BHPL is not specifically designed for HFR, it obtains comparable results with the state-of-the-art. Specifically, BHPL achieves the highest 98.8% accuracy on BUAA NIR-VIS, and 97.3% accuracy close to PACH on CASIA NIR-VIS v2.0. The promising performance of BHPL attributes to three perspectives: 1) the encoder Lightened CNN based on Max-Feature-Map is naturally adaptive to different appearances from different domains [52]; 2) the identity discriminators D^{id} and $\tilde{D}^{id}$ force the learned identity prototype features to encode the identity information accurately; 3) the minimization of the MMD divergence reduces the VIS-NIR domain discrepancy.

Fig. 6. Learned ID card prototypes by BHPL from three random camera photos on NJU-ID. Figures from left to right are: the input camera photos, the learned ID card prototypes, and the groundtruth ID card prototypes.

Fig. 7. Learned camera prototypes by BHPL from three random ID card photos on NJU-ID. Figures from left to right are: the input ID card photos, the learned camera prototypes, and the reference camera photos.

C. Evaluation on ID-camera Dataset

With the popularization of face authentication systems in universities and companies, there is an increasing need to recognize/verify a degraded ID photo stored in the IC card against higher-quality photos captured by digital cameras. In this subsection, we therefore evaluate our BHPL on a representative ID-camera NJU-ID dataset. Note that NJU-ID is quite challenging for HPL due to the extreme lack of within-identity samples. In the case, we treat camera photo domain as Domain A, and ID card domain as Domain B.

Fig. 8. Learned sketch prototypes by BHPL from three random photographs on CUFSF. Figures from left to right are: the input photographs, the learned sketch prototypes, and the groundtruth sketch prototypes.

Fig. 9. Learned photograph prototypes by BHPL from three random sketches on CUFSF. Figures from left to right are: the input sketches, the learned photograph prototypes, and the groundtruth photograph prototypes.

Firstly, we evaluate the learned heterogeneous prototypes by BHPL across camera-to-ID and ID-to-camera domains. Fig. 6 and Fig. 7 illustrate the learned ID card prototypes from three random camera photos, and the learned camera prototypes from three random ID card photos, respectively. Besides, we present the groundtruth ID card photo prototypes in Fig. 6, and the reference camera photos [4] in Fig. 7. It can be observed that our BHPL not only can generate proper ID card prototypes to be stored in the IC card based on contaminated camera photos, but also can restore the degraded low-quality ID card photos into the higher-quality camera prototypes with richer facial details. Although there still exist a few artifacts in the two types of learned heterogeneous prototypes, the identity characteristics are well preserved across domains.

Secondly, we evaluate the learned identity prototype features by BHPL for HFR. Following the setting in [64], we use the camera photos from each identity in the testing set to form the enrolment set, and the ID card photos in the testing set for querying. As suggested in [64] and [65], we choose 3 face synthesis-based methods, i.e., markov weight fields (MWF) [18], Pix2Pix [23], and random sampling with locality constraint (RSLCR) [20], 2 metric learning-based methods, i.e., asymmetric joint learning-based HFR (AJL-HFR) [63] and data augmentation-based joint learning (DA-JL) [64], and 1 feature learning-based method, i.e., coupled attribute learning-based HFR (CAL-HFR) [65], for comparison, and report the rank-50 recognition accuracies in Table V. It can be seen that BHPL achieves the highest recognition accuracy among the compared methods and even delivers 8.7% improvement over the second best CAL-HFR. The inspiring results again verify the effectiveness of the identity prototype feature learning in BHPL. Besides, compared to the metric learning-based and feature learning-based methods, the face synthesis-based methods are less effective because their performance are limited by the small-scale representation set on NJU-ID.

D. Evaluation on Photograph-sketch Face Dataset

Photograph-to-sketch and sketch-to-photograph prototype learning can facilitate the applications of digital entertainment and criminal identification. In this subsection, we therefore evaluate our proposed BHPL on a typical photograph-sketch CUFSF dataset. In the case, we treat photograph domain as Domain A, and sketch domain as Domain B.

[4]It is worth mentioning that, some identities in NJU-ID may not have groundtruth camera photo prototypes.

TABLE VI
RECOGNITION ACCURACIES (%) OF BHPL AND THE OTHER COMPARED METHODS ON CUFSF DATASET.

Methods		Accuracy (%)
Reconstruction-based	MWF [18]	74.2
	SSD [35]	70.9
	RSLCR [20]	75.9
Deep learning-based	FCN [37]	69.8
GAN-based	Pix2Pix [23]	71.4
	MDAL [39]	67.1
Ours	BHPL	**84.5**

As described in Subsection V-A, we adopt two settings for evaluation. Firstly, we adopt the first setting to explore the BHPL's capability for generating sketch prototypes from photographs as well as photograph prototypes from sketches. Secondly, in order to evaluate the learned identity prototype feature by BHPL for HFR, we follow the setting in the existing photograph-sketch synthesis-based methods [20, 39] to make a fair comparison. In this setting, we use the photographs of all testing identities to construct the enrolment set while the corresponding sketches as the query set.

Firstly, we illustrate the learned sketch prototypes from three random photographs in Fig. 8, and the learned photograph prototypes from three random sketches in Fig. 9, respectively. Besides, we also present the groundtruth sketch prototypes or photograph prototypes of the selected identities for reference. From Fig. 8 and Fig. 9, it can be seen that BHPL can generate variation-free sketch prototypes from the input photographs as well as photograph prototypes from the input sketches, although the artistic styles of the two types of images are totally different. In addition, most of these learned sketch (or photograph) prototypes look like the groundtruth sketch (or photograph) prototypes and maintain some crucial personal characteristics such as face contour and eye shape of the input photographs (or sketches). Moreover, we notice that the quality of the generated photograph prototypes from sketches is not as good as the generated sketch prototypes from photographs. This is because that the input sketches usually contain less facial information compared to the input photographs.

Secondly, for the photograph-sketch HFR experiment, we select 6 representative photograph-sketch synthesis-based methods including 3 reconstruction-based methods, i.e., MWF [18], spatial sketch denoising (SSD) [35], and RSLCR [20], 1 deep learning-based method, i.e., FCN [37], and 2 GAN-based methods, i.e., Pix2Pix [23] and MDAL [39], for comparison. Accordingly, we randomly choose 250 identities containing 250 photograph-sketch pairs in the training set to form the representation dictionary for the reconstruction-based MWF, SSD and RSLCR. Moreover, for all the synthesis-based methods, we adopt the rest 300 identities containing 300 synthesized sketches and corresponding groundtruth sketches in the training set to train the Null-space linear discriminant analysis

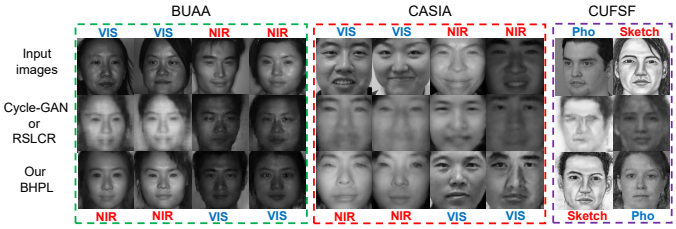


Fig. 10. Comparison between the learned heterogeneous prototypes by BHPL and the synthesized heterogeneous images by two advanced HFS methods, i.e., cycle-GAN and RSLCR, on BUAA NIR-VIS, CASIA NIR-VIS v2.0, and CUFSF datasets. Figures from the top to bottom rows are: the input images, the synthesized images by cycle-GAN (or RSLCR), and the learned prototypes by BHPL.

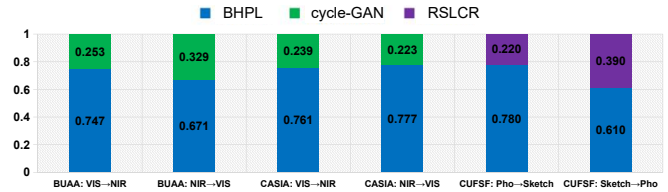


Fig. 11. The preference results of six cross-domain cases on BUAA NIR-VIS, CASIA NIR-VIS v2.0, and CUFSF datasets. The number in each histogram indicate the percentages of preference on the comparison pairs of the learned prototypes by BHPL and the synthesized images by cycle-GAN (or RSLCR).

(NLDA) [66] classifier. We report the rank-1 recognition accuracies of BHPL and these compared methods on CUFSF in Table VI, where we observe that BHPL can further boost the recognition accuracy compared to the other synthesis-based methods with NLDA classifier. This indicates that the learned identity prototype features by BHPL could encode more accurate identity information than the synthesized sketches or photographs, and again demonstrates the rationality of the joint prototype and feature learning in our framework. In addition, we observe that the recognition accuracies of the GAN-based Pix2Pix and MDAL are not competitive with that of RSLCR, although the two GAN-based methods have been shown to synthesize more stylistic sketches or photographs in the texture manner according to [20, 39]. The plausible reason is that Pix2Pix and MDAL based on GAN still produce deformation in the synthesized sketches or photographs as there is no constraint on the local structure, while RSLCR imposes an effective locality constraint [67] that preserves the local reconstruction structure.

E. Comparison with Advanced HFS Approaches

In this subsection, to better reveal the difference between the newly defined HPL and the classic HFS problems, we make a comparison between the learned heterogeneous prototypes by BHPL and the synthesized heterogeneous images by the existing advanced HFS methods on the above three NIR-VIS and photograph-sketch datasets. Specifically, for BUAA NIR-VIS and CASIA NIR-VIS v2.0 datasets, we adopt a representative GAN-based HFS method, i.e., cycle-GAN [22], for comparison; while for the rest photograph-sketch CUFSF dataset, we use the well-known reconstruction-based RSLCR [20] for comparison. We randomly choose 10 query samples from the testing sets of the three evaluated datasets, and then illustrate the corresponding learned cross-domain prototypes by BHPL and the synthesized images by cycle-GAN (or RSLCR) in Fig. 10. In the following, we conclude the key observations and give the analysis:

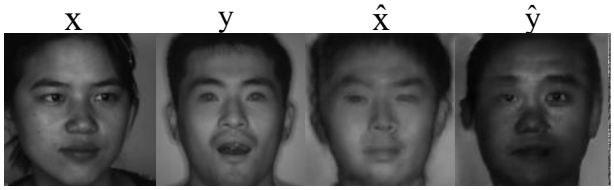
TABLE VII
RECOGNITION ACCURACIES (%) OF BHPL W/O id, BHPL W/O gan, AND BHPL W/O MMD ON BUAA NIR-VIS DATASET.

Methods	Accuracy (%)
BHPL w/o id	58.6
BHPL w/o gan	91.4
BHPL w/o MMD	91.7

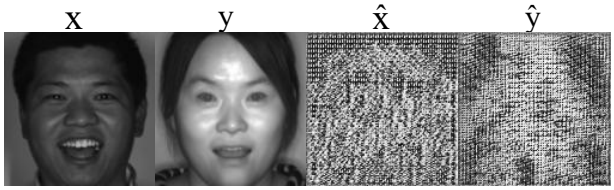
1) The learned cross-domain prototypes by BHPL are standardized and contain almost no variations. By contrast, the facial variations (e.g., expressions and poses) cannot be effectively removed from the synthesized images by cycle-GAN or by RSLCR. For example, on BUAA NIR-VIS and CUFSF, the synthesized VIS (or NIR) images by cycle-GAN and the synthesized sketch by RSLCR still contain pose variations; while on CASIA NIR-VIS v2.0, the mouth areas of the synthesized VIS (or NIR) images by cycle-GAN from the input images with facial expressions are blurred.
2) The learned cross-domain prototypes by BHPL generally have better image quality (e.g., contain fewer artifacts) than the synthesized images by cycle-GAN on the two NIR-VIS datasets. The plausible reason is that, unlike cycle-GAN and the other GAN-based HFS methods trying to approximate the target distribution of face data containing diverse variations, BHPL aims to reconstruct the face prototypes by suppressing the variations, which could alleviate the overfitting to the variations.
3) Although the reconstruction-based RSLCR can synthesize images that capture well the local facial details, e.g., hair style and fringe, there still exists serious distortion in the synthesized sketch from the photograph containing pose variation. This is because the representation dictionary on CUFSF lacks the photograph-sketch pairs possessing the corresponding variations.

Furthermore, we quantitatively compare the quality of the learned cross-domain prototypes by BHPL and the synthesized images by cycle-GAN (or RSLCR) by conducting a user study to ask volunteers to select results that are closer to the real face prototypes through pairwise comparisons. The preference results of six cross-domain cases on the above three datasets are shown in Fig. 11, where we observe that BHPL achieves higher preference scores than that of cycle-GAN and RSLCR in call cases we have tried. This implies that the learned prototypes by BHPL can be easier to be accurately recognized by humans with naked eyes compared to the direct-translated images by cycle-GAN or RSLCR.

Overall speaking, the face synthesis and prototype learning results in the experiments show that the existing GAN-based and reconstruction-based HFS methods may be unsuitable for addressing the new HPL problem, and demonstrate the superiority of our proposed BHPL for HPL from both qualitative and quantitative perspectives.



(a) BHPL w/o id



(b) HPL-GAN w/o gan

Fig. 12. Examples of the learned heterogeneous prototypes by (a) BHPL w/o id and (b) BHPL w/o gan on BUAA NIR-VIS dataset. Figures from left to right are: the input VIS sample $\mathbf{x}$, the input NIR sample $\mathbf{y}$, the learned NIR prototype $\hat{\mathbf{x}}$ for $\mathbf{x}$, and the learned VIS prototype $\hat{\mathbf{y}}$ for $\mathbf{y}$.

F. Ablation Study

In BHPL, there are two types of sub-discriminators, namely identity-related sub-discriminator ($D^{id}\&\tilde{D}^{id}$) and GAN-related sub-discriminator ($D^{gan}\&\tilde{D}^{gan}$). Accordingly, we construct two variants of BHPL, i.e., BHPL w/o id and BHPL w/o gan, by removing $D^{id}\&\tilde{D}^{id}$ and $D^{gan}\&\tilde{D}^{gan}$, respectively, and then study their roles in prototype learning. Fig. 12 illustrates two examples of the learned heterogeneous prototypes by the two BHPL variants on BUAA NIR-VIS dataset. We can see that, when removing $D^{id}\&\tilde{D}^{id}$, BHPL w/o id can still generate the NIR (or VIS) prototype although the restored identity is changed; however, when removing $D^{gan}\&\tilde{D}^{gan}$, BHPL w/o gan cannot even generate visually effective face prototype, which implies that the GAN-related sub-discriminator plays a more important role than the identity-related sub-discriminator in prototype learning. Furthermore, we study the roles of the two sub-discriminators and the MMD constraint in identity prototype feature learning. We list the recognition accuracies of BHPL w/o id, BHPL w/o gan, and BHPL w/o MMD on BUAA NIR-VIS in Table VII, where we observe that BHPL w/o id performs worse than that of BHPL w/o gan and BHPL w/o MMD. We point out that we also have the same observations on the other heterogeneous datasets and omit their results for conciseness. The results indicate that the identity-related sub-discriminator plays more important role in identity prototype feature learning compared to the other two components.

VI. Discussion

Generalization ability: BHPL is a generalized prototype learning framework because it treats the domain style information as one type of facial variation. Consequently, as a by-product, the trained BHPL model is also capable of performing homogeneous prototype learning in a single domain (e.g., Domain A→Domain A or Domain B→Domain B), if the domain style of the input image is exactly the target one. Specifically, given a testing image $\mathbf{x}_t$ from Domain A and a testing image $\mathbf{y}_t$ from Domain B, we switch the input order of the two images by feeding $\mathbf{x}_t$ into G_{encB} and $\mathbf{y}_t$ into G_{encA}. Subsequently, we can acquire the homogeneous prototype $\tilde{\mathbf{x}}_t = G_{decB}(P(\mathbf{x}_t), \mathbf{z}_2)$ for $\mathbf{x}_t$ and $\tilde{\mathbf{y}}_t = G_{decA}(P(\mathbf{y}_t), \mathbf{z}_1)$ for $\mathbf{y}_t$, respectively. Fig. 13(a)-(f) illustrate six examples of the learned homogeneous prototypes by BHPL across NIR→NIR, VIS→VIS, camera→camera, ID→ID, photograph→photograph, and sketch→sketch domains, respectively, where we observe that the facial variations can be successfully removed from the corresponding learned homogeneous prototypes. It is worth mentioning that, BHPL can still be trained on a single-domain face dataset to perform homogeneous prototype learning. Under the circumstances, the Domain A and Domain B are the same, and BHPL would degenerate into a specific homogeneous prototype learning approach similar to [31]. In training, we randomly choose an image $\mathbf{x}$ and an image $\mathbf{y}$ from the same training set as the inputs every time to train BHPL. Fig. 14 illustrates the learned homogeneous prototypes of multiple random images from the testing sets of two VIS-based in-the-wild datasets, i.e., labeled faces in the wild (LFW) [68] and celebrities in frontal-profile (CFP) [69]. It can be seen that BHPL still generates visually-appealing homogeneous prototypes for the contaminated input images containing pose and expression variations in the wild.

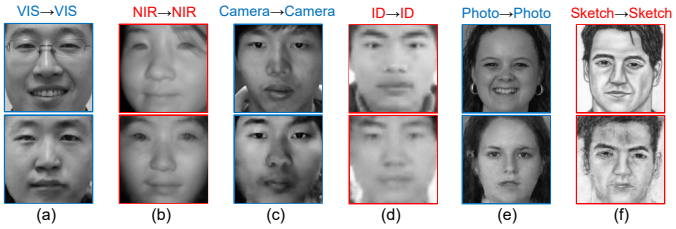


Fig. 13. Examples of the learned homogeneous prototypes by BHPL across NIR→NIR, VIS→VIS, camera→camera, ID→ID, photograph→photograph, and sketch→sketch domains, respectively. Figures from top to bottom rows are the input images and the corresponding learned homogeneous prototypes in the same domain.

Universality: Furthermore, BHPL is designed for disentangling *universal* variations across domains, as it only constrains the identity-preserving and variation-free properties of the learned prototype in the target domain but has no any prior assumption about the variation's type in the source domain. Hence, it is expected that BHPL can be extended to less constrained or even unconstrained scenarios to handle large facial variations (e.g., exaggerated expressions, severe lightings, large poses, and mixed variations). Fig. 15 illustrates the learned heterogeneous prototypes by BHPL from multiple testing samples containing large variations on the less constrained CASIA NIR-VIS v2.0 and NJU-ID datasets, where we observe that BHPL still achieves promising prototypes for a majority of these selected tricky samples on both datasets. Although there is currently no *public in-the-wild* heterogeneous face dataset for evaluating BHPL under totally unconstrained environments across NIR-VIS, camera-card, or photograph-sketch domains, we conjecture that BHPL can potentially handle some unconstrained cases benefiting from its universal design.

Limitations: Although our BHPL has been shown to achieve promising performance for HPL, there still exist two lim-

(a) LFW (b) CFP

Fig. 14. Examples of the learned homogeneous prototypes by BHPL on (a) LFW and (b) CFP datasets. Figures from top to bottom rows are the contaminated input images, the corresponding learned homogeneous prototypes in the same domain, and the reference true prototypes, respectively.

Fig. 15. Heterogeneous Prototype learning results of BHPL under less constrained scenarios. Good examples are in the green box while the relatively bad ones are in the red box. The images from top to bottom lines are: (a) the contaminated input NIR images (or camera photos), (b) the learned VIS prototypes (or card prototypes) by BHPL, and (c) the reference VIS prototypes (or card prototypes).

itations we have not addressed. Firstly, we observe that BHPL sometimes generates nearly the same prototypes for two (or more) similar input samples of different identities. One plausible reason is that the generator in BHPL focuses on generating global consistent prototypes without precisely preserving some crucial local facial characteristics during face synthesis. Secondly, in BHPL's training setting, the training set requires the Domain A and Domain B sets have the same N_d identities. This limits the application of BHPL to the scenarios where the identities of the two domains in the training set are partially overlapped or even totally independent.

VII. Conclusion and Future Works

In this paper, we have focused on a new practical HPL problem, and thus proposed a BHPL framework to tackle it. To the best of our knowledge, BHPL is the first attempt to reconstruct the cross-domain prototype from a contaminated face image using a unified and end-to-end framework. Moreover, BHPL is a joint prototype and feature learning framework that is able to: 1) learn the heterogeneous face prototypes across the source-to-target and target-to-source domains, and 2) learn the discriminative identity prototype features for HFR. Extensive experiments on multiple NIR-VIS, ID-camera, and photograph-sketch heterogeneous face datasets have demonstrated the effectiveness of BHPL.

In the future work, we attempt to impose some locality constraints on the training of the generator, so as to better capture the identity-related local facial characteristics of the input samples in the learned heterogeneous prototypes. Furthermore, we plan to extend BHPL to the more challenging training scenarios where the identities from the source and target domains are partially overlapped or even totally independent, by introducing the cycle-consistency module [22]. Subsequently, for the learned prototypes of the samples from the independent identities, BHPL borrows the cycle-consistency and identity mapping losses from cycle-GAN to *implicitly* preserve the identity; while for the learned prototypes of the samples from the overlapped identities, BHPL still adopts the identity-related discriminator (e.g., D^{id}) to *explicitly* preserve the identity.

References

[1] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution," *Image and Vision Computing*, vol. 56, pp. 28–48, 2016.

[2] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-face: Dual variational generation for heterogeneous face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[3] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[4] X. Pengfei, L. Huang, and C. Liu, "A method for heterogeneous face image synthesis," in *Proceedings of International Conference on Biometrics (ICB)*, 2012, pp. 1–6.

[5] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proceedings of IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018, pp. 83–90.

[6] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1264–1274, 2017.

[7] C. Galea and R. A. Farrugia, "Matching software-generated sketches to face photographs with a very deep cnn, morphed faces, and transfer learning," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421–1431, 2017.

[8] C. Peng, N. Wang, J. Li, and X. Gao, "Face sketch synthesis in the wild via deep patch representation-based probabilistic graphical model," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 172–183, 2019.

[9] S. Duan, Z. Chen, Q. J. Wu, L. Cai, and D. Lu, "Multi-scale gradients self-attention residual learning for face photo-sketch transformation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1218–1230, 2020.

[10] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8913–8925, 2021.

[11] G. Zhang, Y. Chen, W. Lin, A. Chandran, and X. Jing, "Low resolution information also matters: Learning multi-resolution representations for person re-identification," in *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021, pp. 1295–1301.

[12] G. Zhang, J. Yang, Y. Zheng, Y. Wang, Y. Wu, and S. Chen, "Hybrid-attention guided network with multiple resolution features for person re-identification," *Information Sciences*, vol. 578, pp. 525–538, 2021.

[13] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

[14] F. Gao, X. Xu, J. Yu, M. Shang, X. Li, and D. Tao, "Complementary, heterogeneous and adversarial networks for image-to-image translation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3487–3498, 2021.

[15] Y.-C. Chen, X. Xu, and J. Jia, "Domain adaptive image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5274–5283.

[16] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 1005–1010.

[17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 156–163.

[18] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1091–1097.

[19] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 8, pp. 1213–1226, 2012.

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

[21] 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

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6628–6637.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[24] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018.

[25] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7930–7938.

[26] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware cyclegan for face photo-sketch synthesis and recognition," *Pattern Recognition*, vol. 102, p. 107249, 2020.

[27] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020, pp. 800–815.

[28] M. Pang, Y.-M. Cheung, Q. Shi, and M. Li, "Iterative dynamic generic learning for face recognition from a contaminated single-sample per person," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1560–1574, 2020.

[29] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proceedings of the 26th ACM International Conference on Multimedia (ACM MM)*, 2018, pp. 627–635.

[30] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1202–1206.

[31] M. Pang, B. Wang, Y.-m. Cheung, Y. Chen, and B. Wen, "VD-GAN: A unified framework for joint prototype and representation learning from contaminated single sample per person," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2246–2259, 2021.

[32] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[33] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2008.

[34] F. Juefei-Xu, D. K. Pal, and M. Savvides, "Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 141–150.

[35] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014, pp. 800–813.

[36] M. Abdolahnejad and P. X. Liu, "Deep learning for face image synthesis and semantic manipulations: a review and future perspectives," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5847–5880, 2020.

[37] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR)*, 2015, pp. 627–634.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of Advances in Neural Information Processing Systems (NIPs)*, 2014, pp. 2672–2680.

[39] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1419–1428, 2019.

[40] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proceedings of Advances in Neural Information Processing Systems (NIPs)*, 2017, pp. 700–708.

[41] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross spectral face hallucination," in *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019, pp. 1018–1024.

[42] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2545–2560, 2017.

[43] X. Wang and X. Tang, "Bayesian face recognition based on gaussian mixture models," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 4, 2004, pp. 142–145.

[44] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2439–2448.

[45] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[46] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," *Technical Report IRIP-TR-12-FR-001, Beihang University*, vol. 6, 2012.

[47] S. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA nir-vis 2.0 face database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 348–353.

[48] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1814–1826, 2017.

[49] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 513–520.

[50] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[51] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1415–1424.

[52] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[53] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 87–102.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

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

[57] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 451–463, 2016.

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

[59] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *Proceedings of International Conference on Biometrics (ICB)*, 2016, pp. 1–8.

[60] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, no. 1, 2017.

[61] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803–1816, 2019.

[62] M. Cho, T. Kim, I.-J. Kim, K. Lee, and S. Lee, "Relational deep feature learning for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 376–388, 2020.

[63] B. Cao, N. Wang, X. Gao, and J. Li, "Asymmetric joint learning for heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018.

[64] B. Cao, N. Wang, J. Li, and X. Gao, "Data augmentation-based joint learning for heterogeneous face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1731–1743, 2018.

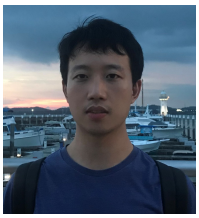
[65] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng, "Coupled attribute learning for heterogeneous face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4699–4712, 2020.

[66] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[67] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3360–3367.
[68] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1978–1990, 2010.
[69] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proceedings of IEEE winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

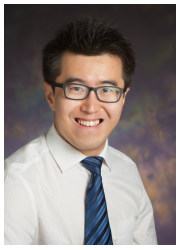
Meng Pang received the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, in 2019, and received the B.Sc. and M.Sc. degrees in software engineering from Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively. He is currently a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include image processing and adversarial machine learning.

Binghui Wang (M'21) is an Assistant Professor of Computer Science at Illinois Institute of Technology, Chicago, USA. He was a postdoctoral researcher in Electrical and Computer Engineering at Duke University from August 2019 to July 2021. He obtained his Ph.D. from the Electrical and Computer Engineering Department at Iowa State University in 2019, and received the M.Sc. and B.Sc. degrees in software engineering and network engineering from Dalian University of Technology, Dalian, China, in 2015 and 2012, respectively. His research interests include adversarial machine learning, data-driven security and privacy, and machine learning. His work has received several honors and awards, including 2020 Amazon Research Award, 2020 DeepMind Best Abstract Award, 2019 NDSS Distinguished Paper Award Honorable Mention, and 2017 INFOCOM Paper for Fast Tracking, etc.



Siyu Huang received the B.E. degree and Ph.D. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2014 and 2019. He was a Visiting Scholar with Language Technologies Institute in School of Computer Science at Carnegie Mellon University in 2018, was a Research Scientist with Big Data Laboratory of Baidu Research from 2019 to 2021, and was a Research Fellow with School of Electrical and Electronic Engineering at Nanyang Technological University. He is currently a postdoctoral fellow with John A. Paulson School of Engineering and Applied Sciences at Harvard University. He has published more than 20 papers on top-tier computer science journals and conferences. His research interests are primarily in computer vision, multimedia analysis, and deep learning.

Yiu-ming Cheung (SM'06-F'18) received the Ph.D. degree from the Department of Computer Science and Engineering at The Chinese University of Hong Kong in Hong Kong. He is an IEEE Fellow, IET Fellow, BCS Fellow, RSA Fellow, and IETI Distinguished Fellow. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China. His research interests include machine learning, pattern recognition, visual computing, and optimization. He serves as an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Pattern Recognition, to name a few.



Bihan Wen (M'14) received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2015 and 2018, respectively. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computational imaging, computer vision, image and video processing, and big data applications. Dr. Wen was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal. He was also a recipient of the Best Paper Runner Up Award at the IEEE International Conference on Multimedia and Expo in 2020. He has been an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology since 2022, and an Associate Editor of MDPI Micromachines since 2021. He has also served as a Guest Editor for IEEE Signal Processing Magazine in 2022.