Heterogeneous Prototype Learning From Contaminated Faces Across Domains via Disentangling Latent Factors

Meng Pang, Binghui Wang, *Member, IEEE*, Mang Ye, *Senior Member, IEEE*, Yiu-Ming Cheung, *Fellow, IEEE*, Yintao Zhou, Wei Huang, and Bihan Wen, *Senior Member, IEEE*

***Abstract*— This article studies an emerging practical problem called heterogeneous prototype learning (HPL). Unlike the conventional heterogeneous face synthesis (HFS) problem that focuses on precisely translating a face image from a source domain to another target one without removing facial variations, HPL aims at learning the variation-free prototype of an image in the target domain while preserving the identity characteristics. HPL is a compounded problem involving two cross-coupled subproblems, that is, domain transfer and prototype learning (PL), thus making most of the existing HFS methods that simply transfer the domain style of images unsuitable for HPL. To tackle HPL, we advocate disentangling the prototype and domain factors in their respective latent feature spaces and then replacing the source domain with the target one for generating a new heterogeneous prototype. In doing so, the two subproblems in HPL can be solved jointly in a unified manner. Based on this, we propose a disentangled HPL framework, dubbed DisHPL, which is composed of one encoder–decoder generator and two discriminators. The generator and discriminators play adversarial games such that the generator embeds contaminated images into a prototype feature space only capturing identity information and a domain-specific feature space, while generating realistic-looking heterogeneous prototypes. Experiments on various heterogeneous datasets with diverse variations validate the superiority of DisHPL.**

Manuscript received 20 May 2022; revised 6 May 2023, 10 November 2023, and 9 January 2024; accepted 18 April 2024. Date of publication 1 May 2024; date of current version 7 April 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62271239 and Grant 62361166629; in part by the NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21; in part by the General Research Fund of RGC under Grant 12201321, Grant 12202622, and Grant 12201323; in part by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02; in part by the Natural Science Foundation of Jiangxi Province under Grant 20232BAB212025; in part by the High-level and Urgently Needed Overseas Talent Programs of Jiangxi Province under Grant 20232BCJ25024; in part by the Jiangxi Double Thousand Plan under Grant JXSQ2023201022; and in part by the Ministry of Education, Republic of Singapore, through its Start-Up Grant and Academic Research Fund Tier 1 under Grant RG61/22. *(Corresponding author: Wei Huang.)*

Meng Pang, Yintao Zhou, and Wei Huang are with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China (e-mail: pangmeng1992@gmail.com; n060101@e.ntu.edu.sg).

Binghui Wang is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: bwang70@iit.edu).

Mang Ye is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: mangye16@gmail.com).

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: ymc@comp.hkbu.edu.hk).

Bihan Wen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: bihan.wen@ntu.edu.sg).

Digital Object Identifier 10.1109/TNNLS.2024.3393072

***Index Terms*— Disentangled representation learning (DRL), generative adversarial learning, heterogeneous face synthesis (HFS), heterogeneous prototype learning (HPL).**

NOMENCLATURE

$\mathbf{x}$	Training image in Domain A, $\mathbf{x} \sim \mathcal{P}_{\text{dataA}}$.
$\mathbf{y}$	Training image in Domain B, $\mathbf{y} \sim \mathcal{P}_{\text{dataB}}$.
$\mathbf{x}^{\text{rp}}$	Real prototype in Domain A, $\mathbf{x}^{\text{rp}} \sim \mathcal{P}_{\text{realA}}$.
$\mathbf{y}^{\text{rp}}$	Real prototype in Domain B, $\mathbf{y}^{\text{rp}} \sim \mathcal{P}_{\text{realB}}$.
$l_x^{\text{id}}/l_x^{\text{var}}$	Identity/variation label of $\mathbf{x}$.
$l_y^{\text{id}}/l_y^{\text{var}}$	Identity/variation label of $\mathbf{y}$.
G	Encoder–decoder generator.
G_{encA}	Encoder A of G.
G_{encB}	Encoder B of G.
$P_{\mathbf{x}}/V_{\mathbf{x}}$	Disentangled prototype/domain feature of $\mathbf{x}$.
$P_{\mathbf{y}}/V_{\mathbf{y}}$	Disentangled prototype/domain feature of $\mathbf{y}$.
G_{dec}	Decoder of G.
$\widehat{\mathbf{x}}/\mathbf{x}^p$	Generated Domain B/A prototype of $\mathbf{x}$.
$\widehat{\mathbf{y}}/\mathbf{y}^p$	Generated Domain A/B prototype of $\mathbf{y}$.
$D, \tilde{D}$	Two multitask discriminators.
$D^{\text{gan}}, \tilde{D}^{\text{gan}}$	Generative adversarial network (GAN)-relevant subdiscriminators in $D, \tilde{D}$.
$D^{\text{id}}, \tilde{D}^{\text{id}}$	Identity-relevant subdiscriminators in $D, \tilde{D}$.
$\mathbf{x}_t$	Query image in Domain A.
$\mathbf{y}_t$	Query image in Domain B.
$\widehat{\mathbf{x}}_t/\mathbf{x}_t^p$	Generated Domain B/A prototype of $\mathbf{x}_t$.
$\widehat{\mathbf{y}}_t/\mathbf{y}_t^p$	Generated Domain A/B prototype of $\mathbf{y}_t$.
$P_{\mathbf{x}_t}/V_{\mathbf{x}_t}$	Disentangled prototype/domain feature of $\mathbf{x}_t$.
$P_{\mathbf{y}_t}/V_{\mathbf{y}_t}$	Disentangled prototype/domain feature of $\mathbf{y}_t$.

I. INTRODUCTION

HETEROGENEOUS face synthesis (HFS) refers to translating a face image from a source domain to another target one through image synthesis. In reality, the domain style can be artistic style (e.g., sketch), light spectrum (e.g., infrared), resolution, and so on. HFS has received increased attention in artificial intelligence (AI) security and can facilitate many applications in law enforcement, criminal identification, person re-identification, digital entertainment, and access control, to name a few [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. In the last decades, a variety of deep generative model-based methods [14], [15], [16], [17], [18] and reconstruction-based methods [19], [20],

[21], [22], [23], [24] have been proposed for addressing HFS. The above-mentioned methods generally hypothesize that the source-domain image is *uncontaminated* and focus on transferring the domain style, for example, from sketch to photograph or from visible (VIS) to near-infrared (NIR), while retaining the facial details unchanged in the target domain.

However, in real-world face retrieval scenarios, a query face image is probably not only inconsistent with the enrollment face image in domain style but also contaminated by diverse facial variations, for example, poses, expressions, misalignments, and disguises/occlusions [25]. Such a combination of domain discrepancy and intrapersonal variance would bring uncertainty to the matching results and easily lead to mismatching. Therefore, it is necessary to simultaneously transfer the domain style and decrease the facial variations of the query face image to reconstruct the variation-free heterogeneous prototype for promoting more accurate matching. This novel and practical problem is defined as *heterogeneous prototype learning* (HPL). Under the circumstances, most existing HFS methods [14], [15], [16], [17], [19], [21], [22], [23], [24], [26] are unsuitable for addressing HPL because these methods only transfer the domain style of images without removing the nuisance facial variations. Moreover, many popular prototype learning (PL)-based approaches [27], [28], [29], [30] are also inapplicable to HPL because these methods concentrate on learning the homogeneous prototypes in the same domain. Based on the above consideration, the motivation of this article is to address HPL by transforming the query face image into the same domain of the enrollment face image and meanwhile decreasing the nuisance facial variations.

Technically, HPL is a compounded problem involving two cross-coupled subproblems, that is, domain transfer and PL. Intuitively, a straightforward idea for addressing HPL is to execute PL and domain transfer sequentially (or vice versa) in a two-step procedure. Nevertheless, it is argued that such a naive two-step solution is unsatisfactory due to its suboptimal design: any image distortion produced in the first step would be magnified when propagating to the next step. Therefore, it is desirable to look for a *one-step* solution to HPL that addresses the above two subproblems *jointly* in a unified framework.

In this article, we thus propose a novel disentangled representation framework, namely **dis**entangled HPL framework (DisHPL), which disentangles and recombines the semantically meaningful prototype and domain factors of a face image for addressing HPL. To be specific, DisHPL hypothesizes that a face image is composed of a prototype, domain, and variation of three main factors, in which the prototype factor is associated with the personal identity information while the domain factor is regarded as a kind of control code that guides the domain direction of prototype generation. Based on this, DisHPL, therefore, advocates disentangling the prototype and source-domain factors from the input face image via disentangled representation and then replacing the source-domain factor with the target domain one to generate the heterogeneous prototype in the target domain, as illustrated in Fig. 1. In doing so, the two subproblems of PL and domain transfer in HPL can be solved jointly in a unified manner. DisHPL is composed of one encoder–decoder generator and two multitask discriminators. Specifically, the generator has two encoders, that is, one encodes the prototype feature and the other encodes the domain-specific feature from a contaminated input image (which could be from the source or the target domain), and one decoder that outputs the homogeneous and heterogeneous prototypes of the input image. The two discriminators contain a GAN-relevant and an identity-relevant subdiscriminators which aim to distinguish between real versus fake prototypes and predict face identity in the source/target domain. The generator competes with the two discriminators to strive for learning: 1) the heterogeneous and homogeneous prototypes of the input image which contain no variations and capture the identity characteristics and 2) the domain-invariant prototype feature of the input image which can be adopted for performing robust heterogeneous face recognition (HFR).

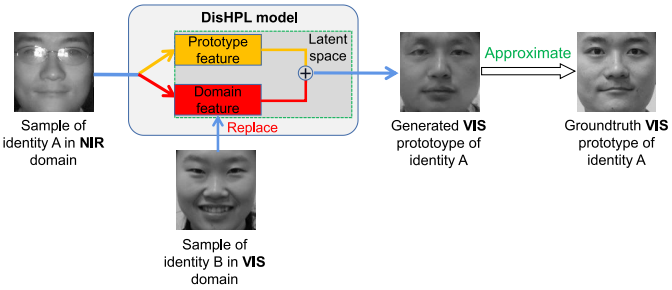


Fig. 1. Illustration of the proposed DisHPL. Given an input sample of identity A wearing glasses from the NIR domain, DisHPL aims to generate its variation-free prototype in the target VIS domain. In DisHPL, the domain and prototype features of the NIR input sample are disentangled in their latent spaces. We replace the domain feature with the target-domain one disentangled from the sample of identity B from the VIS domain and generate the heterogeneous VIS prototype of identity A to approximate the ground-truth VIS one.

It is worth noting that a recent approach [25] has been developed for addressing the HPL problem through bidirectional prototype mapping. The proposed DisHPL can be viewed as an extension of [25], which tackles HPL from a novel perspective of disentangled representation. Compared to [25] specifying a fixed target domain in PL, DisHPL is a more flexible framework that is capable of generating prototypes of different target domains by replacing the source-domain factor with the corresponding target-domain ones. Furthermore, DisHPL is a versatile framework that can learn homogeneous prototypes within the same domain and perform HFR. We summarize the contributions of our work in the following.

1) A novel disentangled representation framework, dubbed DisHPL, is proposed for addressing HPL. It can jointly solve the two subproblems, that is, domain transfer and PL, of HPL in a one-step procedure.
2) An encoder–decoder generator is designed, which simultaneously disentangles the prototype and domain features, and generates the heterogeneous and homogeneous prototypes from a contaminated input image.
3) Two multitask adversarial discriminators are designed, which assist the generator in maintaining the identity

characteristics of the input image and meanwhile removing the facial variations in both generated heterogeneous and homogeneous prototypes.

4) Extensive experiments are conducted on various NIR–VIS and sketch–photograph heterogeneous datasets, which demonstrate the superiority of DisHPL in both tasks of HPL and homogeneous PL, as well as the promising performance for HFR.

The rest of this article is organized as follows. In Section II, we give an overview of the related research works on HFS, PL, and disentangled representation learning (DRL). In Section III, we review the standard GAN. In Section IV, we detail the proposed DisHPL model. Section V evaluates the performance of DisHPL on three NIR–VIS and sketch–photograph heterogeneous datasets from qualitative and quantitative perspectives. In Section VI, we conduct an in-depth discussion of DisHPL in terms of its flexibility, generality, limitations, and comparison with bidirectional HPL (BHPL). Finally, Section VII presents the conclusion and future works.

II. Related Work

A. Heterogeneous Face Synthesis

HFS refers to translating a face image from one domain to another such that heterogeneous images can be evaluated within the same distance space. In the real world, the domain style can be artistic style (e.g., sketch), light spectrum (e.g., infrared), resolution, and so on. Existing HFS methods are roughly classified into two categories [25]: reconstruction-based and deep generative model-based methods.

Reconstruction-based methods [19], [20], [21], [22], [23], [24], [31] usually synthesize the target-domain image based on a learned or predefined source-domain patch dictionary, while preserving the local geometrical structure during face synthesis. For example, Liu et al. [19] adopted the local linear embedding [32] to maintain the local reconstruction structure in the synthesized images. Wang and Tang [20] resorted to a multiscale Markov random field (MRF) to encode smoothness constraints on neighboring sketch patches when transforming a photograph into a sketch. Juefei-Xu et al. [31] synthesized the pseudo-face images across the VIS and NIR domains based on a learned cross-spectral patch dictionary. Wang et al. [24] constructed a simple, yet effective sketch–photograph pair dictionary for sketch synthesis dependent on random sampling and locality-constrained linear coding [33].

Recently, deep generative model-based methods [7], [14], [15], [16], [17], [18] have received wide attention in cross-domain face synthesis. For example, Lezama et al. [34] employed a deep neural network to transfer an NIR image into the VIS domain and then perform a low-rank embedding enhancement. Moreover, benefiting from the powerful generation capability of GAN [35], Isola et al. [15] developed a Pix2Pix package resorting to conditional GAN to translate images across domains. Zhu et al. [14] presented a cycle-consistent GAN (cycle-GAN) to synthesize cross-domain images based on unpaired heterogeneous training data. Fang et al. [26] proposed an identity-aware cycle-GAN (IACycleGAN) framework to employ a new perceptual loss to supervise the image generation network. Liu et al. [36] presented an unsupervised image translation method based on the hypothesis that a pair of heterogeneous images could be mapped into the same representation within a shared-latent space. Zhang et al. [16] developed multidomain adversarial learning (MDAL) to synthesize sketches from photographs by learning the reconstruction procedure from each domain. Song et al. [17] proposed an adversarial discriminative feature learning (ADFL) to combine HFS and feature learning into a joint learning framework. It is worth noting that, these above-mentioned methods treat HFS as a straightforward image translation problem, but cannot effectively decrease variations in the source-domain face images.

Lately, Di et al. [37] developed a domain-agnostic learning-based GAN (DAL-GAN) to synthesize frontal faces in the VIS domain from the thermal faces containing pose variations. Duan et al. [18] presented a pose-agnostic cross-spectral hallucination (PACH) model to handle pose variations by aligning input faces in an unsupervised manner and then performing synthesis based on texture prior information. Note that the above two HFS methods are specifically designed for removing pose variations but cannot generalize to some other facial variations, for example, occlusions and expressions.

B. Prototype Learning

PL is a recent hot topic that targets learning the standard prototype from a contaminated enrollment image containing diverse variations within the same domain. The existing PL methods are roughly classified into two categories [30]: one is to introduce auxiliary sets for image recovery and the other is to train a many-to-one mapping between contaminated images and the prototype.

In the first category, Pang et al. [27] and Gao et al. [38] introduced the unlabeled query set into the labeled enrollment set, thus estimating the prototypes by the clustering centroid of the union of the above two sets through semi-supervised low-rank representation (SSLRR) or Gaussian mixture model (GMM) [39]. In the second category, by virtue of the good mapping capability of GAN, a number of GAN variants [28], [29], [30], [40] are proposed to synthesize realistic-looking prototypes of the contaminated input images while removing the facial variations. For instance, Chen et al. [29] presented an occlusion-aware GAN to detect and restore missing areas in occluded face images. Song et al. [28] proposed a geometry-guided GAN (G2-GAN) by utilizing the fiducial points to guide the facial expression normalization. Huang et al. [40] developed a two-pathway GAN (TP-GAN) to frontalize profile images with poses through local and global transformations. Pang et al. [30] presented a variation disentangling GAN (VD-GAN) to deal with multiple facial variations including expressions, occlusions, and poses. Despite the above-mentioned PL methods having been shown to synthesize high-quality prototypes in a single domain, they are inapplicable to HPL as they do not take the domain style differences into account during the prototype synthesis. More recently, Pang et al. [25] proposed a BHPL framework, which treats the domain discrepancy (e.g., texture difference)

between the source- and target-domain images as a special type of facial variations, thus converting HPL into a generalized PL problem. It is worth noting that, although BHPL can be applied to handle the task of HPL, it is unable to disentangle the domain and prototype factors in the latent space as well and cannot simultaneously generate the heterogeneous and homogenous prototypes for input images as DisHPL did. Furthermore, DisHPL is more flexible in the PL process and could have a broader application prospect compared to BHPL. This is because BHPL specifies a fixed target domain in PL, while DisHPL is capable of generating prototypes of different target domains by replacing the source-domain factor with the corresponding target-domain ones.

C. Disentangled Representation Learning

DRL refers to the learning to factorize the representation of an object (e.g., a face image) into multiple independent representations with each indicating a semantically meaningful factor of the object. Thanks to the variational autoencoder (VAE) [41] and GAN which offer effective tools for extracting disentangled representations, a number of DRL approaches [42], [43], [44], [45], [46] have been developed to disentangle the identity-related feature map and facial attributes from face images for robust face recognition. For example, Liu et al. [42] presented an identity-distilling and dispelling autoencoder to decompose the representation of an image into the identity-distilled and the identity-dispelled components. Tran et al. [43] developed a DRL GAN that rotates the input face image to a specified pose and meanwhile extracts its pose-invariant feature. Zhao et al. [44] proposed an age-invariant model that generates age-invariant features disentangled from the age variations and meanwhile achieves continuous face aging/rejuvenation. Recently, DRL has been widely applied to style transfer, thus promoting many disentangled-based domain transfer approaches [18], [47], [48]. For example, Lee et al. [47] disentangled the representation of a face image into domain-specific and domain-invariant representations to promote learning diverse cross-domain mappings. Inspired by the success of these DRL methods in disentangling different independent factors, we thus introduce DRL into our DisHPL model to disentangle the prototype and domain style factors to perform HPL.

III. Preliminary on GAN

GAN was presented by Goodfellow et al. [35] for training a generative model to synthesize realistic-looking images. GAN is composed of a generator G and a discriminator D, which both can be arbitrary neural networks. The usual way to train D and G is to launch a two-player min-max game. On the one hand, D is trained to classify the fake image generated by G and the real image $\mathbf{x}$. On the other hand, G is trained to generate a realistic-looking image [i.e., $\widehat{\mathbf{x}} = G(\mathbf{z})$] using a random noise $\mathbf{z}$, that aims to fool D. $\mathbf{x}$ is sampled from the data distribution p_{data} (i.e., $\mathbf{x} \sim p_{\text{data}}$), and $\mathbf{z}$ is sampled from the noise distribution p_z (i.e., $\mathbf{z} \sim p_z$). Specifically, the GAN's objective is formulated as follows:

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

According to [35], (1) obtains a global optimal solution while the distribution of these generated images, that is, p_{gen}, is identical to p_{data}. However, in practice, the loss of $\log(1-D(G(\mathbf{z})))$ may saturate because these generated images from G at the beginning of training are poor and can be easily rejected by D. In this case, G is unable to learn anything from zero gradients. To circumvent this issue, Goodfellow et al. advocated using the maximization of $\log(D(G(\mathbf{z})))$ instead of the minimization of $\log(1 - D(G(\mathbf{z})))$, to bring larger gradients early in the learning process. Thus, (1) is rewritten as

$$\max_G V_G = E_{\mathbf{z}}[\log(D(G(\mathbf{z})))] \quad (2)$$

$$\max_D V_D = E_{\mathbf{x}}[\log D(\mathbf{x})] + E_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]. \quad (3)$$

Subsequently, G and D will be updated iteratively by solving (2) and (3) alternatively, until reaching the maximum number of iterations or the convergence condition.

IV. Proposed Model

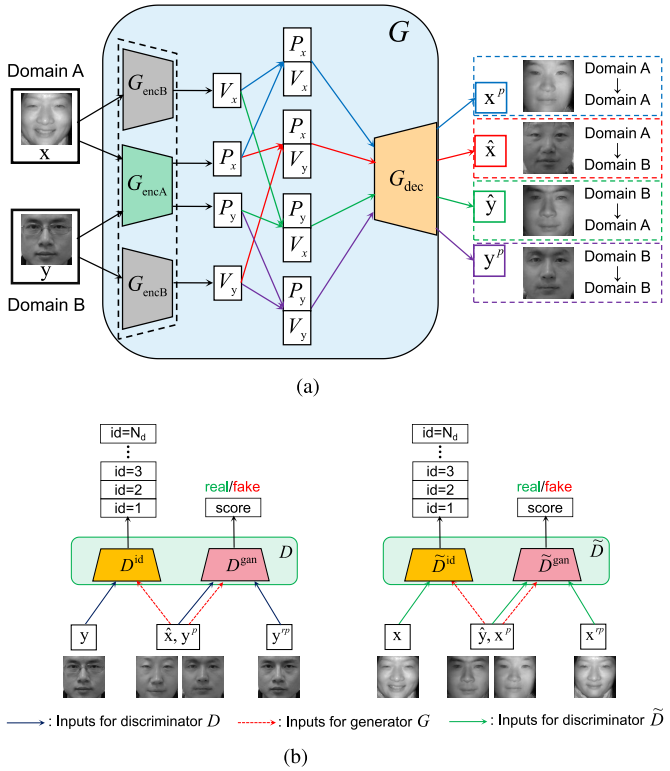
We design our model in this section. We first define the problem and objectives. Next, we describe the architecture of DisHPL, the training algorithm, and the potential applications. The symbols used in DisHPL are summarized in Nomenclature.

A. Problem Definition

We aim to perform: 1) HPL/homogeneous PL in the pixel–spatial space and 2) disentangled feature learning in the latent semantic space from contaminated face images, by using a unified framework.

Suppose we are given a training set that contains images of N_d identities from both Domain A and Domain B. Note that the training set can be unpaired, that is, the images from two domains are not to be one-to-one. In the training set, each image $\mathbf{x}$ in Domain A is annotated with $l_x = \{l_x^{\text{id}}, l_x^{\text{var}}\}$ and is sampled from the data distribution $\mathcal{P}_{\text{dataA}}$, that is, $\mathbf{x} \sim \mathcal{P}_{\text{dataA}}$, while each image $\mathbf{y}$ in Domain B is annotated with $l_y = \{l_y^{\text{id}}, l_y^{\text{var}}\}$ and is sampled from the data distribution $\mathcal{P}_{\text{dataB}}$, that is, $\mathbf{y} \sim \mathcal{P}_{\text{dataB}}$. l_x^{id} (or l_y^{id}) denotes the identity label of $\mathbf{x}$ (or $\mathbf{y}$). l_x^{var} (or l_y^{var}) indicates whether $\mathbf{x}$ (or $\mathbf{y}$) contains variations or not. Take $\mathbf{x}$, for instance, if $\mathbf{x}$ contains arbitrary variation(s) (e.g., pose, expression, and disguise/occlusion), then $l_x^{\text{var}} = 1$; otherwise, $l_x^{\text{var}} = 0$. Next, we select those *uncontaminated* Domain A and Domain B images in the training set by referring to the values of l_x^{var} and l_y^{var}, to build the *real* Domain A and Domain B prototype corpuses, respectively. Each image $\mathbf{x}^{\text{rp}}$ in the real Domain A prototype corpus is sampled from the distribution $\mathcal{P}_{\text{realA}}$, that is, $\mathbf{x}^{\text{rp}} \sim \mathcal{P}_{\text{realA}}$, and each image $\mathbf{y}^{\text{rp}}$ in the real Domain B prototype corpus is sampled from the distribution $\mathcal{P}_{\text{realB}}$, that is, $\mathbf{y}^{\text{rp}} \sim \mathcal{P}_{\text{realB}}$.

Given two random query images $\mathbf{x}_t$ and $\mathbf{y}_t$, one from Domain A and the other from Domain B, DisHPL achieves two objectives as follows.



(a)

(b)

Fig. 2. Illustration of the architecture of DisHPL. (a) G is an encoder–decoder generator for feature disentanglement and prototype generation. (b) D and $\tilde{D}$ are two multitask discriminators for identity classification and prototype distinguishment. $\mathbf{x}$ ($\mathbf{y}$), $\mathbf{x}^{\text{rp}}$ ($\mathbf{y}^{\text{rp}}$), $\widehat{\mathbf{x}}$ (or $\mathbf{x}^p$), and $\widehat{\mathbf{y}}$ (or $\mathbf{y}^p$) denote the input image from Domain A (Domain B), the real Domain A (Domain B) prototype, the learned Domain B (or Domain A) prototype of $\mathbf{x}$, and the learned Domain A (or Domain B) prototype of $\mathbf{y}$, respectively. $P_{\mathbf{x}}$ and $V_{\mathbf{x}}$ ($P_{\mathbf{y}}$ and $V_{\mathbf{y}}$) denote the disentangled prototype and domain features of $\mathbf{x}$ ($\mathbf{y}$), respectively. We use the same two G_{encB} in generator G for the visualization purpose.

1) *HPL:* In the pixel–spatial space, DisHPL aims to reconstruct the Domain B prototype $\widehat{\mathbf{x}}_t$ for $\mathbf{x}_t$ (i.e., from Domain A) and the Domain A prototype $\widehat{\mathbf{y}}_t$ for $\mathbf{y}_t$ (i.e., from Domain B), such that $\widehat{\mathbf{x}}_t$ (or $\widehat{\mathbf{y}}_t$): a) contains no facial variations and b) captures the individual characteristics of $\mathbf{x}_t$ (or $\mathbf{y}_t$).
2) *Disentangled Feature Learning:* In the latent space, DisHPL aims to disentangle: a) the prototype feature $P_{\mathbf{x}_t}$ (or $P_{\mathbf{y}_t}$) that captures the identity information of $\mathbf{x}_t$ (or $\mathbf{y}_t$) and b) the domain feature $V_{\mathbf{x}_t}$ (or $V_{\mathbf{y}_t}$) that contains the Domain A (or Domain B) style information.

As a by-product, DisHPL is capable of performing *homogeneous* PL within the same domain, that is, DisHPL can reconstruct the Domain A prototype $\mathbf{x}_t^p$ for $\mathbf{x}_t$, as well as the Domain B prototype $\mathbf{y}_t^p$ for $\mathbf{y}_t$.

B. DisHPL Architecture

This section introduces the architecture of the proposed DisHPL. As shown in Fig. 2, DisHPL is composed of three major components, that is, one encoder–decoder generator G and two multitask discriminators D and $\tilde{D}$. G competes with D and $\tilde{D}$ to: 1) embed a contaminated image into a prototype feature space and a domain-specific feature space and 2) generate realistic-looking heterogeneous/homogeneous prototypes in the pixel–spatial space.

1) Generator G: G consists of two encoders, that is, G_{encA} and G_{encB}, and one decoder, that is, G_{dec}. G_{encA} encodes the prototype feature $P_{\mathbf{x}}$ of $\mathbf{x}$ and the prototype feature $P_{\mathbf{y}}$ of $\mathbf{y}$, while G_{encB} encodes the domain feature $V_{\mathbf{x}}$ of $\mathbf{x}$ and the domain feature $V_{\mathbf{y}}$ of $\mathbf{y}$. As shown in Fig. 2(a), G_{dec} generates four different prototypes as follows.

1) G_{dec} receives the concatenation of $P_{\mathbf{x}}$ with $V_{\mathbf{x}}$, then generates a *homogeneous* Domain A prototype $\mathbf{x}^p = G_{\text{dec}}(P_{\mathbf{x}}, V_{\mathbf{x}})$ for the Domain A image $\mathbf{x}$ (Domain A→Domain A).
2) G_{dec} receives the concatenation of $P_{\mathbf{x}}$ with $V_{\mathbf{y}}$, then generates a *heterogeneous* Domain B prototype $\widehat{\mathbf{x}} = G_{\text{dec}}(P_{\mathbf{x}}, V_{\mathbf{y}})$ for the Domain A image $\mathbf{x}$ (Domain A→Domain B).
3) G_{dec} receives the concatenation of $P_{\mathbf{y}}$ with $V_{\mathbf{x}}$, then generates a *heterogeneous* Domain A prototype $\widehat{\mathbf{y}} = G_{\text{dec}}(P_{\mathbf{y}}, V_{\mathbf{x}})$ for the Domain B image $\mathbf{y}$ (Domain B→Domain A).
4) G_{dec} receives the concatenation of $P_{\mathbf{y}}$ with $V_{\mathbf{y}}$, then generates a *homogeneous* Domain B prototype $\mathbf{y}^p = G_{\text{dec}}(P_{\mathbf{y}}, V_{\mathbf{y}})$ for the Domain B image $\mathbf{y}$ (Domain B→Domain B).

2) Discriminators D and $\tilde{D}$: As shown in Fig. 2(b), $D = [D^{\text{id}}, D^{\text{gan}}]$ is a multitask discriminator. Specifically, the following holds.

1) D^{id} is an identity-relevant subdiscriminator in D which predicts face identity. It outputs an N_d-dimensional vector in which the location of the maximum value corresponds to the identity label.
2) D^{gan} is a GAN-relevant subdiscriminator that classifies real prototypes and generates fake prototypes by G in Domain B. It gives the real prototype a high score, and the fake one a low score.

In a similar fashion, $\tilde{D} = [\tilde{D}^{\text{id}}, \tilde{D}^{\text{gan}}]$ is still a multitask discriminator including two subdiscriminators. $\tilde{D}^{\text{id}}$ outputs a N_d-dim vector to predict face identity, and $\tilde{D}^{\text{gan}}$ classifies Domain A's real and fake prototypes.

C. DisHPL Training

In DisHPL, there are two adversarial training processes between G and D, and between G and $\tilde{D}$. Accordingly, training DisHPL involves two *alternate* training phases as below.

1) Phase 1: Training of D and G: In this training phase, D and G are trained to compete with each other to force G to generate the heterogeneous prototype $\widehat{\mathbf{x}}$ in the Domain B for the Domain A image $\mathbf{x}$ (Domain A→Domain B), as well as the homogeneous Domain B prototype $\mathbf{y}^p$ for the Domain B image $\mathbf{y}$ (Domain B→Domain B).

For $D = [D^{\text{gan}}, D^{\text{id}}]$, it has two training objectives as follows.

1) Given the generated *fake* Domain B prototypes $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ by G and the *real* Domain B prototype $\mathbf{y}^{\text{rp}}$, D^{gan} targets to classify $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ as two fake prototypes, and $\mathbf{y}^{\text{rp}}$ as the real one.
2) Given the input image $\mathbf{y}$ from Domain B, D^{id} targets to accurately predict the identity label of $\mathbf{y}$, that is, l_y^{id}.

Concretely, the ultimate objective function V_D to train the discriminator D is as follows:

$$\max_D V_D = V_D^{\text{gan}} + \alpha_1 V_D^{\text{id}} \tag{4}$$

where α_1 is a positive balance hyperparameter, and V_D^{gan} and V_D^{id} are denoted as

$$V_D^{\text{gan}} = E_{\mathbf{x}}[\log(1 - D^{\text{gan}}(\widehat{\mathbf{x}}))] + E_{\mathbf{y}}[\log(1 - D^{\text{gan}}(\mathbf{y}^p))] + E_{\mathbf{y}^{\text{rp}}}[\log D^{\text{gan}}(\mathbf{y}^{\text{rp}})] \tag{5}$$

$$V_D^{\text{id}} = E_{\mathbf{y}}[\log D^{\text{id}}_{l_y^{\text{id}}}(\mathbf{y})] \tag{6}$$

where D_i^{id} is the ith element of D^{id}.

The generator G still has two training objectives as follows.

1) Fool D^{gan} to classify both of $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ as two real Domain B prototypes.
2) Force D^{id} to predict the identity label of $\widehat{\mathbf{x}}$ as that of $\mathbf{x}$ (i.e., l_x^{id}) and the label of $\mathbf{y}^p$ as that of $\mathbf{y}$ (i.e., l_y^{id}).

Based on the above-mentioned two objectives, the ultimate objective function V_G to train the generator G is as

$$\max_G V_G = V_G^{\text{gan}} + \lambda_1 V_G^{\text{id}} \tag{7}$$

where λ_1 is a positive balance hyperparameter, and V_G^{gan} and V_G^{id} are denoted as

$$V_G^{\text{gan}} = E_{\mathbf{x},\mathbf{y}}[\log D^{\text{gan}}(\widehat{\mathbf{x}}) + \log D^{\text{gan}}(\mathbf{y}^p)] \tag{8}$$

$$V_G^{\text{id}} * = E_{\mathbf{x},\mathbf{y}}\Big[\log D^{\text{id}}_{l_x^{\text{id}}}(\widehat{\mathbf{x}}) + \log D^{\text{id}}_{l_y^{\text{id}}}(\mathbf{y}^p)\Big]. \tag{9}$$

2) Phase 2: Training of $\tilde{D}$ and G: In this training phase, $\tilde{D}$ and G are trained to compete with each other to force G to generate the heterogeneous prototype $\widehat{\mathbf{y}}$ in Domain A for the Domain B image $\mathbf{y}$ (Domain B→Domain A), as well as the homogeneous Domain A prototype $\mathbf{x}^p$ for the Domain A image $\mathbf{x}$ (Domain A→Domain A).

Similar to D, $\tilde{D} = [\tilde{D}^{\text{gan}}, \tilde{D}^{\text{id}}]$ also has two training objectives as follows.

1) Given the generated *fake* Domain A prototypes $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ by G and the *real* Domain A prototype $\mathbf{x}^{\text{rp}}$, $\tilde{D}^{\text{gan}}$ targets to classify $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ as two fake prototypes, and $\mathbf{x}^{\text{rp}}$ as the real one.
2) Given the input image $\mathbf{x}$ from Domain A, $\tilde{D}^{\text{id}}$ targets to accurately predict the identity label of $\mathbf{x}$, that is, l_x^{id}.

Therefore, the ultimate objective function $\tilde{V}_D$ to train the discriminator $\tilde{D}$ is as follows:

$$\max_{\tilde{D}} V_{\tilde{D}} = V_{\tilde{D}^{\text{gan}}} + \alpha_2 V_{\tilde{D}^{\text{id}}} \tag{10}$$

where α_2 is a positive balance hyperparameter, $V_{\tilde{D}^{\text{gan}}}$ and $V_{\tilde{D}^{\text{id}}}$ are denoted as

$$V_{\tilde{D}^{\text{gan}}} = E_{\mathbf{y}}[\log(1 - \tilde{D}^{\text{gan}}(\widehat{\mathbf{y}}))] + E_{\mathbf{x}}[\log(1 - \tilde{D}^{\text{gan}}(\mathbf{x}^p))] + E_{\mathbf{x}^{\text{rp}}}[\log \tilde{D}^{\text{gan}}(\mathbf{x}^{\text{rp}}) \tag{11}$$

$$V_{\tilde{D}^{\text{id}}} = E_{\mathbf{x}}[\log \tilde{D}^{\text{id}}_{l_x^{\text{id}}}(\mathbf{x})] \tag{12}$$

where $\tilde{D}_i^{\text{id}}$ is the ith entry of $\tilde{D}^{\text{id}}$.

The generator G also has two training objectives as follows.

1) Fool $\tilde{D}^{\text{gan}}$ to classify both of $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ as the real Domain A prototypes.
2) Force $\tilde{D}^{\text{id}}$ to predict the identity of $\widehat{\mathbf{y}}$ as that of $\mathbf{y}$ (i.e., l_y^{id}) and the label of $\mathbf{x}^p$ as that of $\mathbf{x}$ (i.e., l_x^{id}).

In light of the above-mentioned two objectives, the ultimate objective function $\tilde{V}_G$ to train the generator G is as

$$\max_G \tilde{V}_G = \tilde{V}_G^{\text{gan}} + \lambda_2 \tilde{V}_G^{\text{id}} \tag{13}$$

where λ_2 is a positive hyperparameter, $\tilde{V}_G^{\text{gan}}$ and $\tilde{V}_G^{\text{id}}$ are denoted as

$$\tilde{V}_G^{\text{gan}} = E_{\mathbf{x},\mathbf{y}}[\log \tilde{D}^{\text{gan}}(\widehat{\mathbf{y}}) + \log \tilde{D}^{\text{gan}}(\mathbf{x}^p)] \tag{14}$$

$$\tilde{V}_G^{\text{id}} = E_{\mathbf{x},\mathbf{y}}\Big[\log \tilde{D}^{\text{id}}_{l_y^{\text{id}}}(\widehat{\mathbf{y}}) + \log \tilde{D}^{\text{id}}_{l_x^{\text{id}}}(\mathbf{x}^p)\Big]. \tag{15}$$

3) Alternate Training of Phases 1 and 2: To update the two discriminators D and $\tilde{D}$ and the generator G, we alternate run the training of Phases 1 and 2. For clarity, we summarize the alternate training process of DisHPL in Algorithm 1. During the alternate training, the following holds.

1) With D^{gan} and $\tilde{D}^{\text{gan}}$ becoming increasing powerful in classifying real and fake prototypes, G makes an effort to generate the realistic-looking Domain B prototypes $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ to fool D^{gan}, as well as the Domain A prototypes $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ to fool $\tilde{D}^{\text{gan}}$.
2) With D^{id} and $\tilde{D}^{\text{id}}$ being more powerful in predicting face identity, they force the generated $\widehat{\mathbf{x}}$ and $\mathbf{x}^p$ to capture the identity characteristics of $\mathbf{x}$, and the generated $\widehat{\mathbf{y}}$ and $\mathbf{y}^p$ to capture the identity characteristics of $\mathbf{y}$.
3) Moreover, D^{id} and $\tilde{D}^{\text{id}}$ force G_{encA} to encode as much identity information as possible in the learned discriminative prototype features $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$.

Algorithm 1 DisHPL's Training

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

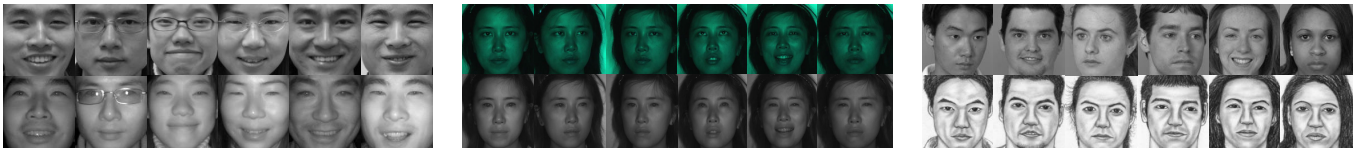
repeat
  Phase 1: Fix G and solve Eqn. (4) to update D
  Phase 1: Fix D and solve Eqn. (7) to update G
  Phase 2: Fix G and solve Eqn. (10) to update $\tilde{D}$
  Phase 2: Fix $\tilde{D}$ and solve Eqn. (13) to update G
until the predefined maximum #iterations is reached or convergence is achieved

Ensure: G, D, $\tilde{D}$

D. DisHPL Applications

After training, we can employ the trained generator G in DisHPL to disentangle the prototype and domain features as well as generate heterogeneous/homogeneous prototypes. Accordingly, DisHPL can handle the following three applications.

1) *HPL:* Generating the Domain A (or Domain B) prototype for a Domain B (or Domain A) image across heterogeneous domains.



(a) (b) (c)

Fig. 3. Illustration of some face examples on (a) CASIA NIR-VIS v2.0, (b) BUAA NIR-VIS, and (c) CUFSF three heterogeneous face datasets.

2) *Homogeneous PL:* Generating the Domain A (or Domain B) prototype for a Domain A (or Domain B) image within the same domain.
3) *HFR:* Given a Domain A (or Domain B) enrollment set and a new Domain B (or Domain A) query image, we could acquire their identity-relevant prototype features in the latent spaces and then perform feature classification.

V. Experimental Results

Section V-A introduces three heterogeneous face datasets to be evaluated followed by the implementation details and parameter settings of DisHPL. In the following, we conduct four experiments to evaluate DisHPL.

1) *Section V-B:* We both qualitatively and quantitatively evaluate the learned heterogeneous/homogeneous prototypes by DisHPL on two NIR–VIS face datasets, that is, CASIA NIR-VIS v2.0 and BUAA NIR-VIS, and on one sketch–photograph face dataset CUFSF.
2) *Section V-C:* We evaluate the proposed DisHPL for HFR on the three heterogeneous face datasets.
3) *Section V-D:* We compare the differences between the synthesized heterogeneous images by two representative HFS techniques, that is, random sampling with locality constraint (RSLCR) and cycle-GAN, and the learned heterogeneous prototypes by DisHPL on the three heterogeneous face datasets.
4) *Section V-E:* We conduct an ablation study to explore the importance of the identity-relevant and GAN-relevant two subdiscriminators on the performance of DisHPL in terms of PL.

A. Experimental Setting

1) Dataset Description: The CASIA NIR–VIS v2.0 [49] is the current largest public and most challenging NIR–VIS face dataset. It contains 725 identities with each having 1–22 VIS and 5–50 NIR images. These involved face images contain diverse facial variations including pose, expression, lighting, and disguises/occlusions (wearing glasses). In the experiment, we adopt the standard evaluation protocol [50] to choose about 6100 NIR and 2500 VIS images of 360 identities as the training set and another 358 identities for testing.

The BUAA NIR–VIS dataset [51] is commonly used for heterogeneous face evaluations across imaging sensors. It includes 150 identities with each possessing nine NIR images and nine VIS images. Each identity's nine NIR (or VIS) images imply nine different variations, that is, happiness, sorrow, anger, surprise, neutral-frontal, tilt-down, tilt-up, right-rotation, and left-rotation. In the experiment, we randomly choose a total of 900 images from 50 identities for training, while the rest 1800 images from 100 identities for testing.

The CUFSF dataset [52] is a commonly used viewed sketch–photograph dataset, which contains 1194 photographs from the FERET dataset [53] and the corresponding 1194 sketches were drawn by the artist. In the experiment, we use two different settings for evaluating the learned heterogeneous/homogeneous prototype and the learned prototype feature, respectively. In the first setting, we randomly select 200 identities for evaluation and then borrow the images with expressions and poses from the five subsets of FERET, namely "bd," "bf," "bg," "bj," and "bk," to expand the photograph size, thus making each identity have one sketch and six photographs. Subsequently, the first 100 identities are used for training, and the remaining 100 are used for testing. In the second setting, we follow the standard evaluation setting in [24] to choose 550 photograph–sketch pairs from 550 identities at random as the training set, and the remaining 644 identities are used for testing.

The above three datasets are publicly available and can only be used for academic research purposes according to the release agreements. Example face images on CASIA NIR–VIS v2.0, BUAA NIR–VIS, and CUFSF datasets are shown in Fig. 3. All face images in these datasets are converted to gray images and center cropped to 128 × 128 pixels.

2) Implementation Details: For the encoder G_{encA}, the Lightened convolutional neural network (CNN) [54] pretrained on the MS-Celeb-1M dataset [55] is employed as the backbone for prototype feature extraction. For the encoder G_{encB}, a different deep neural network, that is, CASIA-Net [56], is adopted as the backbone for extracting the domain features. G_{encA} encodes a 256-D prototype feature while G_{encB} encodes a 50-D domain feature. For the decoder G_{dec}, it takes a 306-D feature vector as the input and outputs a face image of 128 × 128 pixels. For the discriminators D and $\tilde{D}$, they have the same network structure whose input is a face image of 128 × 128 pixels while the output is an $(N_d + 1)$-dim feature vector. The networks of G_{dec} and D (or $\tilde{D}$) are presented in Table I. It is worth mentioning that, the design of the network structures for G, and D (or $\tilde{D}$) in the DisHPL model is flexible, that is, one can customize the design when handling the specific tasks.

We optimize DisHPL by using the stochastic gradient descent with a mini-batch size of 5 and initialize weights of the DisHPL network from a 0-centered Gaussian distribution with a standard deviation of 0.02. As suggested in [56], we adopt Adam [57] as the optimizer, in which the learning rate and momentum values are set to be 0.0002 and 0.5, respectively.

3) Parameter Setting: N_d denotes the total number of the training identities. The four balance hyperparameters, that is, α_1 in (4), λ_1 in (7), α_2 in (10), and λ_2 in (13), are tuned via

TABLE I

NETWORKS OF G_{dec} AND D (OR $\tilde{D}$)

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

TABLE II

PARAMETER SETTINGS OF DISHPL

Datasets	#Train identity	#Test identity	N_d	Balance parameter
CASIA NIR-VIS v2.0	360	358	360	
BUAA NIR-VIS	50	100	50	α_1=α_2=2
CUFSF (setting 1)	100	100	100	λ_1=λ_2=2
CUFSF (setting 2)	550	644	550	

grid search. Empirically, we notice that our DisHPL obtains promising performance when these parameters are all set to be 2 and fix the value across the three evaluated datasets. In Table II, we summarize the parameters' settings on each dataset for clarity.

B. Evaluation of DisHPL for Prototype Learning

This section evaluates the learned heterogeneous and homogeneous prototypes by DisHPL on two NIR–VIS face datasets, that is, CASIA NIR–VIS v2.0 and BUAA NIR–VIS, and a sketch–photograph CUFSF dataset. In the following, we qualitatively and quantitatively evaluate the quality of these learned prototypes.

1) Qualitative Analysis Results: In the first, we *qualitatively* measure the learned prototypes by DisHPL on the three datasets. On the two NIR–VIS datasets, we treat the NIR domain as Domain A, and the VIS domain as Domain B. On the CUFSF dataset, we treat the sketch domain as Domain A, and the photograph domain as Domain B. Given a random query image from Domain A, that is, $\mathbf{x}$, and a random query image from Domain B, that is, $\mathbf{y}$, DisHPL can generate four different prototypes: 1) the homogeneous prototype of $\mathbf{x}$ in Domain A, that is, $\mathbf{x}^p$; 2) the heterogeneous prototype of $\mathbf{x}$ in Domain B, that is, $\widehat{\mathbf{x}}$; 3) the heterogeneous prototype of $\mathbf{y}$ in Domain A, that is, $\widehat{\mathbf{y}}$; and 4) the homogeneous prototype of $\mathbf{y}$ in Domain B, that is, $\mathbf{y}^p$. In Fig. 4, we illustrate six random PL examples of DisHPL on the above three datasets. It can be observed that the following holds.

1) DisHPL successfully learns the *variation-free* heterogeneous prototypes across the VIS-to-NIR, NIR-to-VIS, photograph-to-sketch, and sketch-to-photograph domains, as well as the homogeneous prototypes within the same VIS, NIR, sketch, and photograph domains. Intuitively, for these contaminated input images containing variations of different facial expressions (e.g., happiness and surprise), occlusions of different types of glasses, slight left/right head posture, and misalignment, DisHPL is capable of simultaneously transferring the domain styles and removing the facial variations.
2) From a visual perspective, most of the learned heterogeneous and homogeneous prototypes by DisHPL preserve the personal identity characteristics of the contaminated input images well and look similar to the reference ground-truth prototypes.
3) There exist a few artifacts and slight deformations on the generated photograph (or sketch) prototypes on CUFSF, which leads to a lower image quality compared to that of the generated VIS (or NIR) prototypes on the other two NIR–VIS datasets. This is because the artistic styles of the sketch and photographs are very different, and the only decoder in DisHPL is not amenable to producing two images of such different styles at the same time. In addition, the image quality of the generated photograph prototypes from sketches seems not as good as that of the generated sketch prototypes from photographs. The plausible reason could be that the input sketches generally provide very few facial details compared to the input photographs during prototype generation.

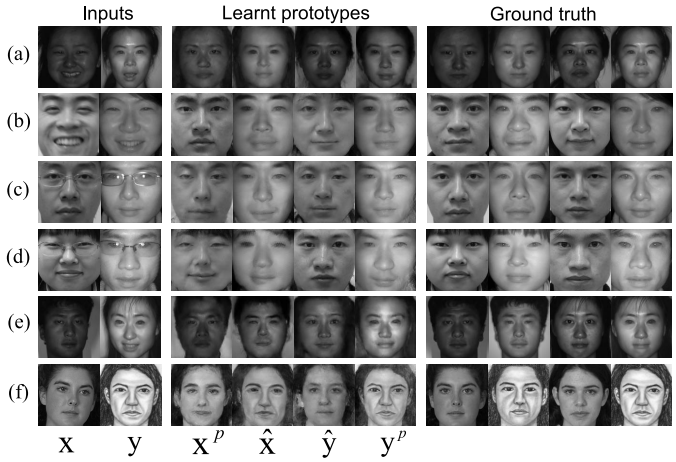


Fig. 4. Six randomly selected PL examples (a)–(f) of DisHPL on BUAA NIR–VIS, CASIA NIR–VIS v2.0, and CUFSF datasets. In each row, figures from left to right are: the input query image $\mathbf{x}$ and $\mathbf{y}$ from two different domains, the four learned prototypes by DisHPL, that is, $\mathbf{x}^p$ (the learned homogeneous prototype of $\mathbf{x}$), $\widehat{\mathbf{x}}$ (the learned heterogeneous prototype of $\mathbf{x}$), $\widehat{\mathbf{y}}$ (the learned heterogeneous prototype of $\mathbf{y}$), and $\mathbf{y}^p$ (the learned homogeneous prototype of $\mathbf{y}$), and the corresponding ground-truth prototypes for reference.

2) Quantitative Analysis Results: Note that many existing evaluation metrics such as L_1/L_2 distance or structural similarity index metric (SSIM) [58] are actually designed to measure the pixel/structure similarity between synthesized image and the input, while DisHPL has altered the facial structure (e.g., frontalizing a face with large pose) during the face synthesis. Therefore, we follow the work in [47] and conduct a user study that asks volunteers to artificially judge the quality of the learned prototypes by DisHPL through pairwise verification. In each dataset, we sample multiple pairs of query images randomly in the source domain and their

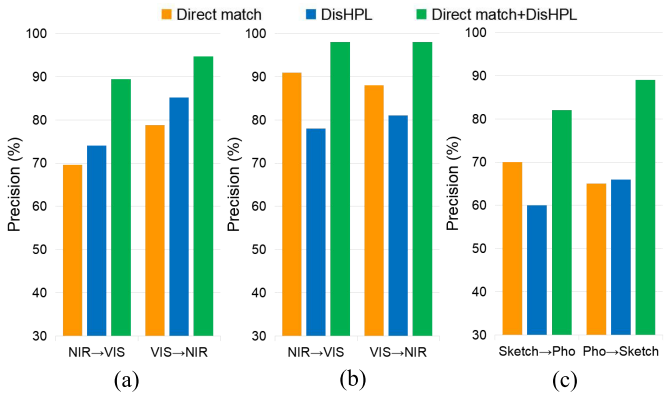


Fig. 5. Average precisions (%) of three different verification strategies, that is, "Direct match," DisHPL, and "Direct match + DisHPL," on (a) CASIA NIR–VIS v2.0, (b) BUAA NIR–VIS, and (c) CUFSF datasets.

ground-truth prototypes in the target domain and then use the learned prototype of each query image to match with the ground-truth prototype for verification. For reference, we also directly use the original query images for matching and denote the strategy as "Direct match." Furthermore, we take both of the above two strategies into consideration and propose a new verification strategy denoted as "Direct match + DisHPL." Specifically, we assume the identities of the pair of query images and ground-truth prototype match if either of the two strategies gives a matching decision. The purpose of "Direct match + DisHPL" is to test whether the learned prototypes by DisHPL can assist volunteers in recognizing the contaminated query images from different domains (e.g., NIR images or hand-drawn sketches) more accurately. The average precisions of different strategies are shown in Fig. 5.

As shown in Fig. 5, DisHPL achieves relatively good precisions (60%–85%) in the six cases we have tried across the three datasets, which indicates that the majority of the learned prototypes by DisHPL capture the personal identity characteristics well. Besides, the precisions of DisHPL also exceed that of the "Direct match" in half the cases. Furthermore, we observe that the fusion strategy of "Direct match + DisHPL" significantly improves the verification performance compared to either DisHPL or "Direct match," which shows the complementarity of the above two verification strategies as well as demonstrates that the learned prototypes by DisHPL could assist verification/recognition in practice.

C. Evaluation of DisHPL for HFR

This section evaluates the HFR performance of the proposed DisHPL on CASIA NIR–VIS v2.0, BUAA NIR–VIS, and CUFSF datasets. For DisHPL, we acquire the identity prototype features for both VIS (or photograph) enrollment and NIR (or sketch) query images through the encoder of the trained DisHPL model and then use a cosine similarity-based nearest neighbor (NN) classifier to conduct prediction.

In the NIR–VIS HFR experiment, we select **12** NIR–VIS feature learning-based methods, involving four *handcrafted feature learning-based* kernelized discriminative spectral regression (KDSR) [59], kernelized margin-based cross-modality metric learning (KMCM2L) [60], common encoding feature discriminant (CEFD) [61], and hierarchical hyperlingual-words local binary pattern (H2-LBP3) [62], and eight *deep-learning-based* transfer NIR–VIS HFR network (TRIVET) [50], domain-specific units nets (DSU-Nets) [63], ADFL [17], relational deep feature learning (RGM) [64], channel augmented joint learning (CAJL) [65], heterogeneous face interpretable disentangled representation (HFIDR) [7], invariant deep representation (IDR) [66], and PACH [18]. Table III shows the rank-1 recognition rates of DisHPL and the comparing feature learning-based methods on the two NIR–VIS datasets.

TABLE III
RECOGNITION RATES (%) OF DISHPL AND THE NIR–VIS FEATURE LEARNING-BASED METHODS ON CASIA NIR–VIS V2.0 AND BUAA NIR–VIS FACE DATASETS

Methods		CASIA NIR-VIS v2.0	BUAA NIR-VIS
Hand-crafted	KDSR [59]	37.5	83.0
	KMCM2L [60]	76.0	–
	CEFD [61]	85.6	–
	H2-LBP3 [62]	43.8	88.8
Deep learning	TRIVET [50]	95.7	93.9
	DSU-Nets [63]	96.3	–
	ADFL [17]	98.2	95.2
	RGM [64]	97.2	97.6
	CAJL [65]	–	98.3
	IDR [66]	97.3	94.3
	PACH [18]	**98.9**	98.6
	HFIDR [7]	98.6	–
Ours	DisHPL	97.5	**98.7**

TABLE IV
RECOGNITION RATES (%) OF DISHPL AND THE OTHER SKETCH–PHOTOGRAPH SYNTHESIS-BASED METHODS ON THE CUFSF DATASET

Methods		Accuracy (%)
Reconstruction-based	SSD [67]	70.9
	MWF [22]	74.2
	RSLCR [24]	75.9
GAN-based	Pix2Pix [15]	71.4
	MDAL [16]	67.1
	IACycleGAN [26]	74.5
	eaeSPADE [68]	70.2
	SCG [69]	78.0
Ours	DisHPL	**84.5**

In sketch–photograph HFR experiment, we use eight sketch–photograph synthesis-based methods for comparison. They are three *reconstruction-based* methods including spatial sketch denoising (SSD) [67], Markov weight fields (MWF) [22], and RSLCR [24], and five *GAN-based* methods including Pix2Pix [15], MDAL [16], IACycleGAN [26], edge-aware enhancement spatially adaptive denormalization (eaeSPADE) [68], and semi-cycle-GAN (SCG) [69]. In accordance with [24], 250 sketch–photograph pairs from 250 random identities in the training set are used to construct the representation dictionary for the three reconstruction-based methods. Subsequently, the remaining 300 identities containing 300 synthesized sketches and the ground-truth sketches are used for training a null-space linear discriminant analysis (NLDA) [70] classifier for all the synthesis-based methods to perform HFR. Table IV lists the rank-1 recognition rates of DisHPL and the comparing sketch–photograph synthesis-based methods on the CUFSF dataset.

From Tables III and IV, we can observe that the following holds.

1) Although DisHPL is not specifically designed for HFR, it still achieves promising performance in both

NIR–VIS and sketch–photograph HFR tasks. For example, DisHPL performs the best on the CUFSF and BUAA NIR-VIS datasets and obtains slightly inferior results compared to the state-of-the-art PACH on the CASIA NIR–VIS v2.0 dataset. This indicates that the learned prototype features by DisHPL capture well the identity information across domains. The promising performance of DisHPL in HFR owes to its two major advantages: 1) the Max-Feature-Map-based Lightened CNN in G_{encA} is adaptive to different appearances in different modalities [54] and 2) the two identity-relevant subdiscriminators, that is, D^{id} and $\tilde{D}^{\text{id}}$, trained in the source and target domains *explicitly* preserve the identity in the learned prototype features across domains.

2) On the two NIR–VIS datasets, the deep-learning-based methods usually obtain higher recognition rates compared to the conventional handcrafted feature learning-based methods, which verifies the good representation learning capability of deep neural networks in HFR.
3) PACH achieves an improvement of 1.4% over our proposed DisHPL on the CASIA NIR–VIS v2.0 dataset, while the results are comparable on the BUAA NIR–VIS dataset. The effectiveness of PACH in HFR lies in its design of an unsupervised face alignment (UFA) module to align the facial shape of the input NIR or VIS image, thus enabling more accurate extraction of identity information. However, it needs to be noted that, unlike DisHPL which can directly generate face prototypes in an end-to-end manner, PACH requires first executing UFA and then texture prior synthesis (TPS). Furthermore, since the facial shape in the UFA module generally only considers pose variations or misalignment, it cannot handle some other facial variations like expressions and disguises/occlusions, as DisHPL can.
4) On the sketch–photograph CUFSF dataset, the recognition rates of all synthesis-based methods with the trained NLDA classifier are not competitive with that of DisHPL (prototype feature + NN classifier). The results imply that the learned prototype features by DisHPL may be more suitable for HFR than the synthesized sketches (or photographs) by these synthesis-based methods, thus verifying the rationality of the joint feature and PL in DisHPL. Besides, we are interested to find that while Pix2Pix, MDAL, and eaeSPADE synthesize sketches/photographs with more stylization than the reconstruction-based RSLCR in terms of texture, the three GAN-based methods perform worse than RSLCR for HFR. The possible reason lies in that GAN-based MDAL, Pix2Pix, and eaeSPADE would produce distortions when synthesizing sketches/photos because they lack restrictions on local structures, which would adversely affect identity preservation. In contrast, RSLCR introduces an effective local constraint using local linear coding (LLC) [33], which can maintain local geometric structures. SCG achieves better HFR results compared to RSLCR and other GAN-based methods. This is because it constructs pseudo-photo-sketch pairs to supervise the generators and introduces a cycle-consistent loss with noise injection to mitigate the steganography effect during photo or sketch synthesis.

D. Comparison With Representative HFS Approaches

HPL and HFS handle different applications: HFS focuses on translating images across domains, while HPL aims to preserve personal identity as well as remove the existing facial variations during domain transferring. To better illustrate the differences between the classic HFS and new HPL problems, this section compares the differences between the synthesized heterogeneous images via the representative HFS methods and learned heterogeneous prototypes by DisHPL on the three NIR–VIS and sketch–photograph face datasets. Accordingly, on CASIA NIV–VIS v2.0 and BUAA NIR–VIS, we choose the typical GAN-based Cycle-GAN [14] for comparison; while on CUFSF, we adopt the popular reconstruction-based RSLCR [24], which is specifically designed for sketch–photograph synthesis, as the comparing method. In Fig. 6(a), we illustrate the synthesized images by cycle-GAN or RSLCR and the cross-domain prototypes by DisHPL from six random contaminated input images on the three datasets. The key observations and the corresponding analysis are as follows.

1) For cycle-GAN or RSLCR, they are still unable to effectively remove the facial variations, for example, poses and expressions, in their synthesized heterogeneous images. By contrast, the learned cross-domain prototypes by DisHPL are standardized and contain almost no variations, which are visually similar to the ground truth face prototypes.
2) On the two NIR–VIS datasets, DisHPL usually generates higher-quality learned cross-domain prototypes with fewer artifacts and noises, compared to the synthesized heterogeneous images by cycle-GAN. This is because, cycle-GAN and many other existing GAN-based HFS methods usually try to approximate the target distribution of the original face data including diverse facial variations, which may easily cause over-fitting limited by the small-scale training data. By contrast, our DisHPL targets to approximate a shrunken distribution of standard face prototypes, which could mitigate the over-fitting of the variations.
3) On the sketch–photograph CUFSF dataset, although the reconstruction-based RSLCR retains the local facial details (e.g., fringe and hair style) to some extent in its synthesized sketch and photograph, we note that there exist serious distortions in the synthesized sketch from the photograph with profile posture. The reason is that the performance of RSLCR depends on the richness of the representation dictionary but the CUFSF dataset cannot provide the corresponding sketch–photograph pairs that possess the pose variations.

Subsequently, we compare the image quality of the synthesized heterogeneous images by cycle-GAN or RSLCR, and the learned cross-domain prototypes by DisHPL from the quantitative perspective. Following the strategy in [47], we perform a user study to invite multiple volunteers to choose

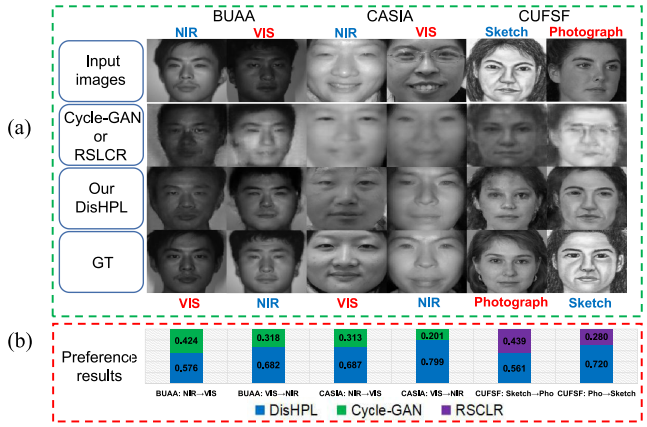


Fig. 6. (a) Comparison between the synthesized heterogeneous images by cycle-GAN or RSLCR and learned heterogeneous prototypes by DisHPL on CASIA NIR–VIS v2.0, BUAA NIR–VIS, and CUFSF. From top to bottom rows: the input images, the synthesized images by cycle-GAN (or RSLCR), the learned prototypes by DisHPL, and the ground-truth prototypes. (b) Preference results that answer the question "which image is more similar to the ground truth?" The number in the bar chart indicates the percentage of preference on that comparison pair.

the synthesized image or learned prototype that represents the identity of the target person more accurately through pairwise comparisons. In Fig. 6(b), we show the preference results of six cross-domain cases on the above three evaluated datasets. It can be seen that DisHPL always achieves higher preference scores than RSLCR and cycle-GAN in all six cases. The inspiring results indicate that, compared to these direct-translated images by RSLCR or cycle-GAN, the learned prototypes by DisHPL could be much more easily recognized by humans. Generally speaking, the qualitative and quantitative comparison results in Fig. 6 indicate that the existing reconstruction-based and GAN-based HFS methods may be unsuitable for tackling the new HPL problem, as well as verify the effectiveness of DisHPL for HPL.

E. Ablation Study

In DisHPL, there exist two subdiscriminators: one is the GAN-relevant subdiscriminator (D^{gan} & $\tilde{D}^{\text{gan}}$) and the other is the identity-relevant subdiscriminator (D^{id} & $\tilde{D}^{\text{id}}$). This section studies the importance of the two subdiscriminators on the performance of DisHPL in terms of PL. Subsequently, we construct two DisHPL's variants, that is, DisHPL w/o gan and DisHPL w/o id, by removing D^{gan} & $\tilde{D}^{\text{gan}}$ and D^{id} & $\tilde{D}^{\text{id}}$, respectively, and then evaluate their performance in PL.

Figs. 7 and 8 illustrate the learned prototypes by the two DisHPL variants on CASIA NIR–VIS v2.0 and BUAA NIR–VIS datasets, respectively. It can be seen that, when removing D^{gan} & $\tilde{D}^{\text{gan}}$, DisHPL w/o *gan* cannot even generate visually effective homogeneous and heterogeneous prototypes; when removing D^{id} & $\tilde{D}^{\text{id}}$, DisHPL w/o *id* still generates the variation-free prototypes in the correct target domains but fail to preserve the identity. The results demonstrate that the GAN-relevant subdiscriminator is more important compared to the identity-relevant subdiscriminator in homogeneous and HPL. Furthermore, we explore the importance of the aforementioned GAN-relevant and identity-relevant subdiscriminators in identity prototype feature learning. On BUAA

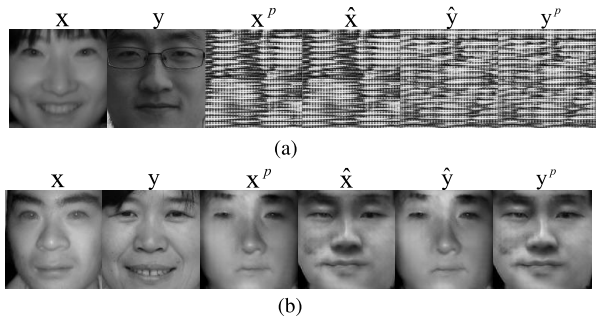


Fig. 7. PL examples of (a) DisHPL w/o gan and (b) DisHPL w/o id on CASIA NIR–VIS v2.0. Left to right columns: the input image $\mathbf{x}$ from the NIR domain, the input image $\mathbf{y}$ from the VIS domain, the learned NIR prototype $\mathbf{x}^p$ of $\mathbf{x}$, the learned VIS prototype $\widehat{\mathbf{x}}$ of $\mathbf{x}$, the learned NIR prototype $\widehat{\mathbf{y}}$ of $\mathbf{y}$, and the learned VIS prototype $\mathbf{y}^p$ of $\mathbf{y}$.

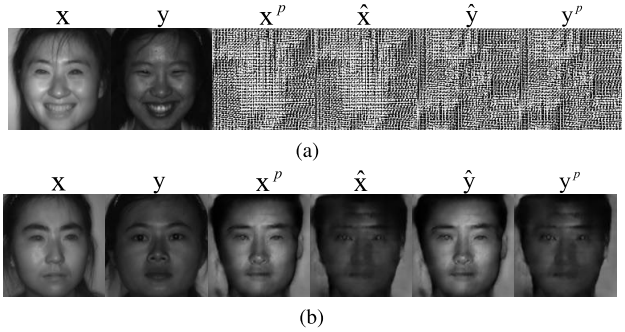


Fig. 8. PL examples of (a) DisHPL w/o gan and (b) DisHPL w/o id on the BUAA NIR-VIS dataset. Left to right columns: the input image $\mathbf{x}$ from the NIR domain, the input image $\mathbf{y}$ from the VIS domain, the learned NIR prototype $\mathbf{x}^p$ of $\mathbf{x}$, the learned VIS prototype $\widehat{\mathbf{x}}$ of $\mathbf{x}$, the learned NIR prototype $\widehat{\mathbf{y}}$ of $\mathbf{y}$, and the learned VIS prototype $\mathbf{y}^p$ of $\mathbf{y}$.

NIR–VIS, DisHPL w/o gan achieves a much higher recognition rate (74.6%) than that of DisHPL w/o gan (41.8%), indicating that the identity-relevant subdiscriminator plays a more important role than the GAN-relevant subdiscriminator in identity prototype feature learning.

VI. Discussion

A. Flexibility of DisHPL

Given a contaminated face image from a source domain (e.g., Domain A), DisHPL can reconstruct its homogeneous prototype in the same domain as well as the heterogeneous prototype in the target domain (e.g., Domain B). When generating the heterogeneous prototype, it is flexible for DisHPL to replace the source-domain feature with the target-domain one disentangled from any arbitrary random image in the target domain. Take the CAISA NIR–VIS v2.0 dataset, for example, we randomly choose four VIS images in the training set to disentangle four target-domain features and then illustrate the corresponding learned VIS prototypes by DisHPL from a testing-contaminated NIR image in Fig. 9. From Fig. 9, it can be seen that these four learned VIS prototypes from the NIR testing image look almost the same. *This implies that DisHPL would not leak identity information when disentangling the domain from the prototype.*

B. Generality of DisHPL

Despite some recent methods [18], [37] attempting to perform cross-domain pose frontalization or face alignment

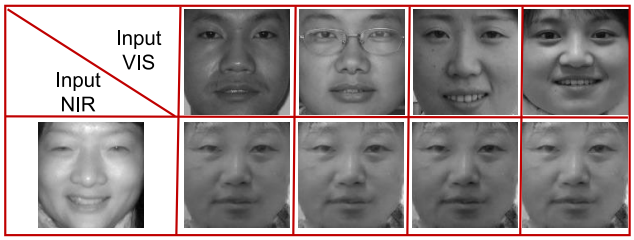


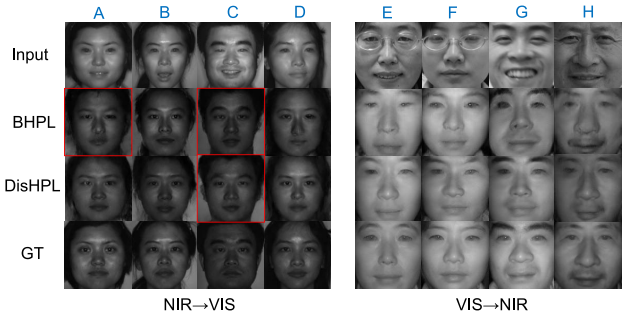
Fig. 9. Comparison of four learned VIS prototypes by DisHPL from an NIR input image on CASIA NIR–VIS v2.0 dataset. Figures in the first low are the four randomly chosen VIS images that provide the target-domain features. Figures in the second row are the NIR input image and the corresponding four learned VIS prototypes.

during face synthesis, they are ad hoc methods specifically designed to handle pose variations or misalignments but unable to generalize to other variations such as occlusions and expressions. *In contrast, DisHPL is a generic PL framework that can be applied to universal variations, including (but not limited to) expressions, poses, misalignments, and occlusions/disguises.* This is because DisHPL only constrains the variation-free and identity-preserving properties of the learned prototypes but has no prior assumption about the type(s) of the input variations. In fact, DisHPL focuses on learning a straightforward end-to-end *transformation* between contaminated images and the standard prototypes. Therefore, it is expected that our proposed DisHPL can be easily extended to other NIR–VIS and sketch–photograph heterogeneous face datasets containing *diverse* variations.

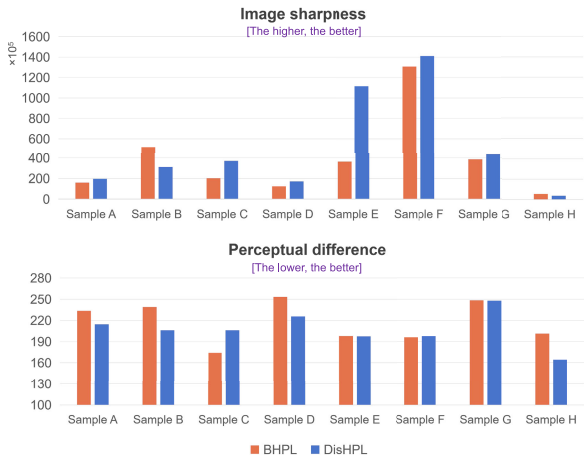
C. Comparison With BHPL

In this section, we compare the performance of our DisHPL with the closely related BHPL on the HPL task from both qualitative and quantitative perspectives. We randomly selected eight input images, four from the NIR domain of the BUAA NIR–VIS dataset and four from the VIS domain of the CASIA NIR–VIS dataset. In Fig. 10(a), we show the target-domain heterogeneous prototypes generated by BHPL and DisHPL, along with the GT prototypes. It can be observed that the majority of the prototypes generated by BHPL and DisHPL are of good quality. Furthermore, from a visual perspective, the generated prototypes by DisHPL appear to be of better quality and closer to the GT for the input samples A, D, E, and G compared to those generated by BHPL. Additionally, in the cases of input samples B, F, and H, the quality of the prototypes generated by DisHPL and BHPL is comparable.

Subsequently, we employ two analytical-based quantitative metrics, namely *image sharpness* and *perceptual difference*, to measure the image clarity and identity preservation of the generated prototypes, respectively. The image sharpness [71] is measured using the Brenner algorithm, which accumulates the squares of the differences between horizontally neighboring pixels. The perceptual difference [72] is measured by calculating the L_1-distance between the identity features extracted from the learned prototype and its GT prototype within the same domain. Note that a higher value of image sharpness indicates higher image clarity, whereas a smaller value of perceptual difference indicates that the identity of the learned prototype is closer to that of the input. As shown in Fig. 10(b),



(a)



(b)

Fig. 10. (a) Qualitative and (b) quantitative comparison results of BHPL and our DisHPL for HPL. (a) Figures from top to bottom rows are the input images from the NIR (or VIS) domain, the learned VIS (or NIR) prototypes by BHPL, the learned VIS (or NIR) prototypes by DisHPL, and the reference VIS (or NIR) GT prototypes. The figures in red boxes indicate the generated unsatisfactory prototypes. (b) Comparison of the values of the two quantitative metrics, that is, image sharpness and perceptual difference, between BHPL and DisHPL.

out of the 16 metric values for the eight images, 11 metric values favor DisHPL over BHPL, indicating that DisHPL has certain advantages in terms of prototype generation quality and identity preservation. This may be attributed to the successful disentanglement of identity-related information from domain-related information in our model. Additionally, we observe that the quantitative comparison results in Fig. 10(b) align closely with the qualitative analysis results in Fig. 10(a), thus validating the rationality and effectiveness of the aforementioned two quantitative metrics.

In addition to the slight improvement in HPL performance, we emphasize that DisHPL has two important advantages over BHPL. First, DisHPL is a powerful framework that can simultaneously perform HPL across domains and homogeneous PL within the same domain. Second, DisHPL is more flexible than BHPL because it decouples prototype and domain factors. This enables DisHPL to potentially handle prototype reconstruction in different target domains by replacing the source-domain factor with different target-domain ones.

D. Limitations of DisHPL

Despite DisHPL has shown to obtain promising heterogeneous and homogeneous prototypes across the VIS-to-NIR,

NIR-to-VIS, sketch-to-photograph, and photograph-to-sketch domains, DisHPL still suffers from three following limitations.

First, through the experiments, it is observed that DisHPL usually generates almost the same prototypes for a few similar input images with different identities, especially on BUAA NIR–VIS and CUFSF datasets. The plausible reasons are twofold: 1) the training sets of the two evaluated datasets are small,[1] which are unable to provide sufficient identity label information for training the identity-relevant discriminators in DisHPL and can easily lead to over-fitting problem and 2) DisHPL concentrates on learning global consistent prototypes without constraints on the local structure. Under the circumstances, some local facial characteristics that can distinguish two similar persons may be lost during the prototype generation. To alleviate the over-fitting problem caused by the small training data, one possible solution is to generate massive paired heterogeneous images by virtue of some powerful generative models, for example, DVG-Face [8], and then screen out high-quality ones to augment the training set. Furthermore, to enhance identity preservation during prototype generation, it may be feasible to introduce the attention mechanism [73] into our DisHPL model to better capture the local facial characteristics with identity-distinguishing effects.

Second, the proposed DisHPL relies on a single decoder, which may not be amenable to generating paired heterogeneous images with very different artistic styles simultaneously. As a result, the generated prototypes by DisHPL may have low image quality in specific image transformation situations (e.g., the transformation between the photograph and the sketch). To tackle this issue, it may be feasible to design a more flexible decoder by adding the adaptive instance normalization (AdaIN) [74] layers which adapt the decoder to arbitrary domain styles by changing the affine parameters of the layers accordingly, or just replacing the single decoder with dual decoders in DisHPL to generate images of different styles, respectively.

Third, in practice, it is difficult to achieve arbitrary target-domain prototype reconstruction with DisHPL. This is because it requires a prerequisite of having samples with multiple domains in the training set to learn different domain factors for replacing the source domain. However, existing publicly available datasets typically only have two domains (such as NIR–VIS and photograph–sketch). To address this issue, we also provide a potentially viable solution by further improving the DisHPL model into a multidirectional mapping model similar to StarGAN [75], followed by training on multiple bidomain datasets and utilizing mask vector to control the model's learning of target-domain factors from different datasets. We will leave the interesting study as the future research directions.

VII. Conclusion and Future Works

This article has studied an emerging challenging HPL problem, which involves two coupled subproblems of domain transfer and PL. To tackle HPL, we have proposed the DisHPL to jointly address the above two subproblems in a unified disentangled representation framework. Given a contaminated face image from the source domain, DisHPL can simultaneously: 1) disentangle its domain and prototype features and 2) generate proper heterogeneous and homogeneous prototypes. Empirically studies on various NIR–VIS and sketch–photograph datasets have verified the superiority of DisHPL in both HPL and HFR tasks.

In the future study, we plan to impose the local attention module on DisHPL to better preserve the local facial characteristics with identity-distinguishing effects during PL. Besides, we attempt to develop a more flexible decoder by introducing the AdaIN layers to enable DisHPL to simultaneously generate higher-quality paired prototypes with very different artistic styles. Moreover, inspired by the success of DisHPL in disentangling prototype and domain factors, we will try to further disentangle the variation factor for facial variation manipulation, thus extending the proposed DisHPL to the interesting application of heterogeneous face editing/interpolation.

[1]Limited by the acquisition condition, the current public heterogeneous face datasets are small and usually contain dozens or hundreds of identities.

References

[1] X. Pengfei, L. Huang, and C. Liu, "A method for heterogeneous face image synthesis," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 1–6.

[2] 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 Vis. Comput.*, vol. 56, pp. 28–48, Dec. 2016.

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

[4] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo-sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3096–3108, Oct. 2019.

[5] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 83–90.

[6] P. Li, B. Sheng, and C. L. P. Chen, "Face sketch synthesis using regularized broad learning system," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5346–5360, Oct. 2022.

[7] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5611–5625, Oct. 2022.

[8] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-face: Dual variational generation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2938–2952, Jun. 2022.

[9] M. Zhang, N. Wang, Y. Li, and X. Gao, "Neural probabilistic graphical model for face sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2623–2637, Jul. 2020.

[10] M. Pang, Y.-M. Cheung, B. Wang, and R. Liu, "Robust heterogeneous discriminative analysis for face recognition with single sample per person," *Pattern Recognit.*, vol. 89, pp. 91–107, May 2019.

[11] J. Liu et al., "Identity preserving generative adversarial network for cross-domain person re-identification," *IEEE Access*, vol. 7, pp. 114021–114032, 2019.

[12] M. Zhang, N. Wang, Y. Li, and X. Gao, "Deep latent low-rank representation for face sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3109–3123, Jan. 2019.

[13] M. Pang, B. Wang, S. Chen, Y.-M. Cheung, R. Zou, and W. Huang, "Cross-domain prototype learning from contaminated faces via disentangling latent factors," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2022, pp. 4369–4373.

[14] 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.*, Oct. 2017, pp. 2223–2232.

[15] 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.*, Jul. 2017, pp. 1125–1134.

[16] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1419–1428, May 2019.

[17] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 32, no. 1, 2018, pp. 7355–7362.

[18] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7927–7935.

[19] 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.*, Jun. 2005, pp. 1005–1010.

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

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

[22] 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.

[23] 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.

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

[25] M. Pang, B. Wang, S. Huang, Y.-M. Cheung, and B. Wen, "A unified framework for bidirectional prototype learning from contaminated faces across heterogeneous domains," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1544–1557, 2022.

[26] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware CycleGAN for face photo-sketch synthesis and recognition," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107249.

[27] 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 Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1560–1574, Apr. 2021.

[28] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 627–635.

[29] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1202–1206.

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

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

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

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[34] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6807–6816.

[35] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

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

[37] X. Di, S. Hu, and V. M. Patel, "Heterogeneous face frontalization via domain agnostic learning," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 01–08.

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

[39] X. Wang and X. Tang, "Bayesian face recognition based on Gaussian mixture models," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 4, Aug. 2004, pp. 142–145.

[40] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[42] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2080–2089.

[43] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3007–3021, Dec. 2019.

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

[45] S. Zhao, J. Li, and J. Wang, "Disentangled representation learning and residual GAN for age-invariant face verification," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107097.

[46] M. Pang, B. Wang, M. Ye, Y.-M. Cheung, Y. Chen, and B. Wen, "DisP+V: A unified framework for disentangling prototype and variation from single sample per person," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 867–881, Feb. 2023.

[47] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.

[48] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5102–5112.

[49] 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.

[50] 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.

[51] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, 2012, vol. 6.

[52] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 513–520.

[53] 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.

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

[55] 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.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[56] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

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

[58] W. Zhou, 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, pp. 600–612, 2004.

[59] 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.

[60] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, Jun. 2018.

[61] 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.

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

[63] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1803–1816, Jul. 2019.
[64] M. Cho, T. Kim, I.-J. Kim, K. Lee, and S. Lee, "Relational deep feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 376–388, 2021.
[65] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 13567–13576.
[66] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proc. AAAI. Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 2000–2006.
[67] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 800–813.
[68] C. Zhang, D. Liu, C. Peng, N. Wang, and X. Gao, "Edge aware domain transformation for face sketch synthesis," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2761–2770, 2022.
[69] C. Chen, W. Liu, X. Tan, and K.-Y.-K. Wong, "Semi-supervised cycle-GAN for face photo-sketch translation in the wild," *Comput. Vis. Image Understand.*, vol. 235, Oct. 2023, Art. no. 103775.
[70] 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 Recognit.*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.
[71] L. Her and X. Yang, "Research of image sharpness assessment algorithm for autofocus," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2019, pp. 93–98.
[72] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[73] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1972–1987, Apr. 2023.
[74] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.
[75] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

Meng Pang received the B.Sc. and M.Sc. degrees in software engineering from Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, in 2019.

He was a Post-Doctoral Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2020 to 2022. He is currently a Distinguished Professor with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China. His research interests include image processing, artificial intelligence security, and artificial intelligence medical.

Binghui Wang (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Iowa State University, Ames, IA, USA, in July 2019.

He held a post-doctoral position at the Electrical and Computer Engineering Department, Duke University, Durham, NC, USA, from August 2019 to July 2021. He is currently an Assistant Professor of computer science with the Illinois Institute of Technology, Chicago, IL, USA. His research interests include adversarial machine learning, data-driven security and privacy, and machine learning.

Mang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and 2016, respectively, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2019.

He was a Research Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is currently a Full Professor at Wuhan University. His research interests include multimedia retrieval, computer vision, and pattern recognition.

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2000.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China. His research interests include machine learning, pattern recognition, visual computing, and optimization.

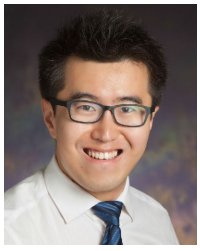
Dr. Cheung is an IET Fellow, a BCS Fellow, an RSA Fellow, and an IETI Distinguished Fellow. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and *Pattern Recognition*, to name a few.

Yintao Zhou received the B.Eng. and M.Eng. degrees in computer science and technology from Nanchang University, Nanchang, China, in 2020 and 2023, respectively, where he is currently pursuing the Ph.D. degree under the supervision of Prof. Wei Huang and Prof. Meng Pang.

His research interests mainly include computer vision and deep learning.

Wei Huang received the B.Eng. and M.Eng. degrees from Harbin Institute of Technology, Harbin, China, in 2004 and 2006, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2010.

He was with the University of California at San Diego, La Jolla, CA, USA, and the Agency for Science Technology and Research, Singapore, as a Post-Doctoral Research Fellow. He is currently a Professor at the Department of Computer Science, Nanchang University, Nanchang, China. His main research interests include AI, machine learning, computer-aided diagnosis, and image processing.



Bihan Wen (Senior Member, IEEE) 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 was then a Researcher with Dolby Laboratories, San Francisco, CA, USA. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests span areas of machine learning, computer vision, image and video processing, computational imaging, and big data applications.

Dr. Wen is currently a member of the IEEE Computational Imaging Technical Committee. He was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal of Singapore.