Cross-domain Prototype Learning from Contaminated Faces via Disentangling Latent Factors

Meng Pang
Nanchang University
pangmeng1992@gmail.com

Binghui Wang
Illinois Institute of Technology
bwang70@iit.edu

Shenbo Chen
Nanchang University
ccb02kingdom@gmail.com

Yiu-ming Cheung
Hong Kong Baptist University
ymc@comp.hkbu.edu.hk

Rong Zou
Hong Kong Baptist University
rongzou@comp.hkbu.edu.hk

Wei Huang*
Nanchang University
n060101@e.ntu.edu.sg

ABSTRACT

This paper focuses on an emerging challenging problem called heterogeneous prototype learning (HPL) across face domains—It aims to learn the *variation-free* target domain prototype for a contaminated input image from the source domain and meanwhile preserve the personal identity. HPL involves two coupled subproblems, i.e., *domain transfer* and *prototype learning*. To address the two subproblems in a unified manner, we advocate disentangling the prototype and domain factors in their respected latent feature spaces, and replace the latent source domain features with the target domain ones to generate the heterogeneous prototype. To this end, we propose a disentangled heterogeneous prototype learning framework, dubbed DisHPL, which consists of one encoder-decoder generator and two discriminators. The generator and discriminators play adversarial games such that the generator learns to embed the contaminated image into a *prototype feature space* only capturing identity information and a *domain-specific feature space*, as well as generating a realistic-looking heterogeneous prototype. The two discriminators aim to predict personal identities and distinguish between real prototypes versus fake generated prototypes in the source/target domain. Experiments on various heterogeneous face datasets validate the effectiveness of DisHPL.

CCS CONCEPTS

• **Computing methodologies → Biometrics**; *Reconstruction*; Image representations.

KEYWORDS

Heterogeneous prototype learning, heterogeneous face recognition, domain transfer, generative adversarial network.

ACM Reference Format:
Meng Pang, Binghui Wang, Shenbo Chen, Yiu-ming Cheung, Rong Zou, and Wei Huang. 2022. Cross-domain Prototype Learning from Contaminated Faces via Disentangling Latent Factors. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557571

1 INTRODUCTION

Heterogeneous face synthesis (HFS) refers to translating a face image from a source domain to another target one through image synthesis. HFS has received increasing attention in many applications such as criminal identification and digital entertainment [19]. A variety of reconstruction-based methods [1, 5, 32, 37] and deep generative model-based methods [4, 11, 16, 38] have been developed for addressing HFS. These methods generally hypothesize that the source domain image is *uncontaminated*; and focus on transferring the domain style, e.g., from near infrared (NIR) to visible (VIS), while retaining the facial details unchanged in the target domain.

However, the source domain face images captured in real world are likely to be contaminated by diverse facial variations, e.g., poses, expressions, or occlusions. Under the circumstances, most existing HFS methods [1, 4, 11, 28, 32, 32, 37, 38] would make the identity of the synthesized target domain image difficult to be recognized by humans as these methods only transfer the image's domain style without decreasing the nuisance facial variations. Consequently, it is critical to reconstruct the *variation-free* face prototype across the source-to-target different domains to better represent the personal identity. This novel practical problem is defined as *heterogeneous prototype learning (HPL)* [21]. Unlike the classic HFS which simply performs image-to-image translation, HPL targets to *simultaneously* preserve the personal identity and remove the facial variations during domain transferring. Therefore, the above-mentioned HFS methods are unsuitable for HPL because they cannot effectively handle the facial variations during face synthesis. Furthermore, we note that the existing prototype learning-based approaches [2, 9, 20, 22, 27] are inapplicable to HPL because they concentrate on learning the homogeneous prototypes within the same domain.

HPL involves two coupled subproblems, i.e., domain transfer and prototype learning. A straightforward idea for addressing HPL is to sequentially execute prototype learning and domain transfer (or vice versa) in a two-step procedure. Nevertheless, we contend that this two-step solution is unsatisfactory due to its sub-optimal design: any image distortion produced in the first step will be magnified when propagating to the second step. Hence, we are motivated to look for a desired solution to HPL that is capable of addressing the above two subproblems *jointly* using a unified framework.

*Wei Huang is the corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557571

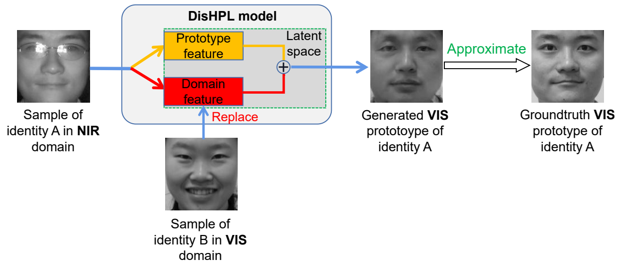

Figure 1: Illustration of DisHPL. Given a sample of identity A wearing glasses in NIR domain, its prototype and domain features are disentangled in their latent spaces. DisHPL replaces the domain feature with the target domain one disentangled from the sample of identity B in VIS domain, and generates the VIS prototype of identity A to approximate the groundtruth VIS one.

Inspired by the success of disentangled representation learning [12, 26, 31] based on generative adversarial network (GAN) [7] for factorizing semantically-meaningful factors of a object, we advocate disentangling the prototype and source domain factors from an input face image, and then replace the source domain factor with the target domain one to generate the heterogeneous prototype across the source-to-target domains, as shown in Figure 1. To this end, we propose a disentangled heterogeneous prototype learning (DisHPL) framework. In DisHPL, the prototype factor is associated with the personal identity information, while the domain factor is treated as a kind of control code that guides the domain direction of prototype generation. DisHPL is composed by one encoder-decoder generator and two multi-task discriminators. Specifically, the generator has two encoders, i.e., one encodes the prototype feature and the other encodes the domain-specific feature from a contaminated input image, and one decoder that outputs both the homogeneous and heterogeneous prototypes of the input image. The two discriminators both contain an identity-relevant and a GAN-relevant sub-discriminators which aim to predict face identity and classify real and fake prototypes in the source/target domain. The generator competes with the two discriminators to force: 1) the learnt heterogeneous and homogeneous prototypes to contain no variations and capture the identity characteristics of the input image; and 2) the learnt prototype feature to accurately encode the identity-relevant information, which could be adopted for performing robust heterogeneous face recognition (HFR).

Contributions: 1) We propose the DisHPL, which is the first attempt to tackle HPL using a unified disentanglement learning framework. **2)** We design an encoder-decoder generator, which simultaneously disentangles the prototype and domain features, and generates the heterogeneous and homogeneous prototypes from a contaminated image; **3)** We perform experiments on various NIR-VIS and sketch-photo heterogeneous datasets, to validate the effectiveness of DisHPL for both heterogeneous and homogeneous prototype learning, and the promising performance for HFR.

2 RELATED WORK

HFS targets to synthesize cross-domain face images and then evaluate them within the same domain. Liu *et al.* [17] firstly explored the HFS issue and adopted the local linear embedding [23] to maintain the local reconstruction structure in the synthesized images. Subsequently, a variety of reconstruction-based methods [1, 5, 29, 32, 37] have been proposed to synthesize face images in the target domain dependent on a learnt or pre-defined source domain patch dictionary. GAN [7] has also received considerable attention in HFS. For instance, Zhu *et al.* [38] presented a cycle-consistent GAN to generate cross-domain images using unpaired heterogeneous training data. Lately, a few attempts [3, 4, 16, 28] have been made to combine HFS and feature learning into a joint learning framework. These above-mentioned methods treat HFS as a straightforward image-to-image translation problem, but cannot effectively remove facial variations existed in the source domain face images.

Prototype learning is a recent hot topic that tries to learn the face prototype from a contaminated enrolment image containing nuisance variations. e.g., poses, expressions, and occlusions. Benefiting from GAN's powerful mapping capability, a number of GAN variants [2, 9, 13, 20, 27] have been proposed to remove facial variations and to generate the realistic-looking prototypes. For instance, Song *et al.* [27] put forward a geometry-guided GAN to perform expression normalization by utilizing fiducial points. Huang *et al.* [9] developed a two-pathway GAN to frontalize profile images with poses through local and global transformations. Despite promising *homogeneous* prototypes obtained by these methods, they are unable to learn *heterogeneous* prototypes across domains.

3 THE PROPOSED MODEL

3.1 Problem Definition

Suppose a training set contains N_d identities from both Domain A and Domain B. Each image $\mathbf{x}$ in Domain A is sampled from the distribution $\mathcal{P}_{dataA}$, i.e., $\mathbf{x} \sim \mathcal{P}_{dataA}$, and is annotated with $l_x = \{l_x^{id}, l_x^{var}\}$; while each image $\mathbf{y}$ in Domain B is sampled from the distribution $\mathcal{P}_{dataB}$, i.e., $\mathbf{y} \sim \mathcal{P}_{dataB}$, and is annotated with $l_y = \{l_y^{id}, l_y^{var}\}$. 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 facial variations or not. Take $\mathbf{x}$ for example, if $\mathbf{x}$ contains arbitrary variation(s) (e.g., pose, expression, and occlusion/disguise), then $l_x^{var} = 1$; otherwise $l_x^{var} = 0$. 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 in the real Domain A prototype corpus is denoted as $\mathbf{x}^{rp} \sim \mathcal{P}_{realA}$, and each image in the real Domain B prototype corpus is denoted as $\mathbf{y}^{rp} \sim \mathcal{P}_{realB}$.

Given two query images $\mathbf{x}_t$ and $\mathbf{y}_t$, one from Domain A and the other from Domain B, DisHPL has two objectives:

- **Heterogeneous prototype learning**: Learning an appropriate Domain B prototype $\widehat{\mathbf{x}}_t$ for $\mathbf{x}_t$ (that is from Domain A) and Domain A prototype $\widehat{\mathbf{y}}_t$ for $\mathbf{y}_t$ (that is from Domain B), such that $\widehat{\mathbf{x}}_t$ (or $\widehat{\mathbf{y}}_t$): 1) contains no facial variations, and 2) captures the identity characteristics of $\mathbf{x}_t$ (or $\mathbf{y}_t$).
- **Disentangled feature learning**: Disentangling 1) 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 2) the domain feature $V_{\mathbf{x}_t}$ (or $V_{\mathbf{y}_t}$) that contains Domain A (or Domain B) specific information.

As a by-product, DisHPL is also able to perform *homogeneous* prototype learning within the same domain.

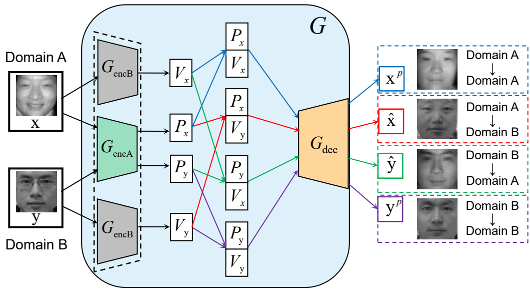

Figure 2: Illustration of the generator in DisHPL. $\mathbf{x}$ ($\mathbf{y}$), $\widehat{\mathbf{x}}$ (or $\mathbf{x}^p$), and $\widehat{\mathbf{y}}$ (or $\mathbf{y}^p$) denote the input image from Domain A (Domain B), the learnt Domain B (or Domain A) prototype of x, and the learnt Domain A (or Domain B) prototype of y, respectively. $P_\mathbf{x}$ and $V_\mathbf{x}$ ($P_\mathbf{y}$ and $V_\mathbf{y}$) denote the disentangled prototype and domain features of x (y), respectively.

3.2 DisHPL Architecture

3.2.1 Generator G. G consists of two encoders, i.e., G_{encA} and G_{encB}, and one decoder, i.e., G_{dec}. G_{encA} encodes the prototype feature $P_\mathbf{x}$ for $\mathbf{x}$ and the prototype feature $P_\mathbf{y}$ for $\mathbf{y}$; while G_{encB} encodes the domain feature $V_\mathbf{x}$ for $\mathbf{x}$ and the domain feature $V_\mathbf{y}$ for $\mathbf{y}$. Subsequently, G_{dec} receives the concatenation of $P_\mathbf{x}$ and $V_\mathbf{x}$, the concatenation of $P_\mathbf{x}$ and $V_\mathbf{y}$, the concatenation of $P_\mathbf{y}$ and $V_\mathbf{x}$, the concatenation of $P_\mathbf{y}$ and $V_\mathbf{y}$ as four inputs, and then generates four different prototypes, i.e., $\mathbf{x}^p$, $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$, and $\mathbf{y}^p$, as illustrated in Figure 2.

3.2.2 Discriminators D and $\tilde{D}$. D has two sub-discriminators D^{id} and D^{gan}. D^{id} is an identity-relevant sub-discriminator for predicting the face identity in Domain B. It outputs a vector of N_d-dimension with N_d indicating the number of identities in the training set. D^{gan} is a GAN-relevant sub-discriminator for classifying real and fake prototypes in Domain B. It gives a high score to the real prototype and a low score to the fake one. Similarly, $\tilde{D}$ is also a multi-task discriminator involving two sub-discriminators $\tilde{D}^{id}$ and $\tilde{D}^{gan}$. $\tilde{D}^{id}$ outputs a N_d-dimensional vector for face identity prediction in Domain A, while $\tilde{D}^{gan}$ is adopted to classify real and fake prototypes in Domain A.

3.3 DisHPL Training

DisHPL involves two adversarial training processes between G and D, and between G and $\tilde{D}$. Accordingly, we propose to use two *alternate* phases to train DisHPL.

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

For $D = [D^{gan}, D^{id}]$, it has **two** training objectives: **1)** Given the generated *fake* Domain B prototypes $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ by G and the *real* Domain B prototype $\mathbf{y}^{rp}$, D^{gan} wishes to classify $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ as two fake prototypes, and meanwhile classify $\mathbf{y}^{rp}$ as the real one. **2)** Given the Domain B input image $\mathbf{y}$, D^{id} wishes to predict its identity label l_y^{id} correctly. Hence, the ultimate objective function V_D to train the discriminator D is as

$$\max_D V_D = V_D^{gan} + \alpha_1 V_D^{id}, \tag{1}$$

where α_1 is a balance parameter. V_D^{gan} and V_D^{id} are defined as $V_D^{gan} = E_{\mathbf{y}^{rp}}[\log D^{gan}(\mathbf{y}^{rp})] + E_\mathbf{x}[\log(1 - D^{gan}(\widehat{\mathbf{x}}))] + E_\mathbf{y}[\log(1 - D^{gan}(\mathbf{y}^p))]$ and $V_D^{id} = E_\mathbf{y}[\log D^{id}_{l_y^{id}}(\mathbf{y})]$, where D_i^{id} is the i-th element in D^{id}.

For G, it also has **two** training objectives: **1)** Fool D^{gan} to classify both of $\widehat{\mathbf{x}}$ and $\mathbf{y}^p$ as the real Domain B prototypes; **2)** Enable D^{id} to predict the identity label of $\widehat{\mathbf{x}}$ as that of $\mathbf{x}$ (i.e., l_x^{id}), and the identity label of $\mathbf{y}^p$ as that of $\mathbf{y}$ (i.e., l_y^{id}). Hence, the ultimate objective function V_G to train the generator G is as

$$\max_G V_G = V_G^{gan} + \lambda_1 V_G^{id}, \tag{2}$$

where λ_1 is a balance parameter. V_G^{gan} and V_G^{id} are defined as $V_G^{gan} = E_{\mathbf{x},\mathbf{y}}[\log D^{gan}(\widehat{\mathbf{x}}) + \log D^{gan}(\mathbf{y}^p)]$ and $V_G^{id} = E_{\mathbf{x},\mathbf{y}}[\log D^{id}_{l_x^{id}}(\widehat{\mathbf{x}}) + \log D^{id}_{l_y^{id}}(\mathbf{y}^p)]$, respectively.

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

For $\tilde{D} = [\tilde{D}^{gan}, \tilde{D}^{id}]$, it has **two** training objectives similar to D: **1)** Given the generated *fake* Domain A prototypes $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ by G and the *real* Domain A prototype $\mathbf{x}^{rp}$, $\tilde{D}^{gan}$ wishes to classify $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ as two fake prototypes, and meanwhile classify $\mathbf{x}^{rp}$ as the real one. **2)** Given the Domain A input image $\mathbf{x}$, $\tilde{D}^{id}$ wishes to predict its identity label l_x^{id} accurately. Therefore, the ultimate objective function $\tilde{V}_D$ to train the discriminator $\tilde{D}$ is as follows:

$$\max_{\tilde{D}} V_{\tilde{D}} = V_{\tilde{D}^{gan}} + \alpha_2 V_{\tilde{D}^{id}}, \tag{3}$$

where α_2 is a balance parameter. $V_{\tilde{D}^{gan}}$ and $V_{\tilde{D}^{id}}$ are defined as $V_{\tilde{D}^{gan}} = E_{\mathbf{x}^{rp}}[\log \tilde{D}^{gan}(\mathbf{x}^{rp})] + E_\mathbf{y}[\log(1 - \tilde{D}^{gan}(\widehat{\mathbf{y}}))] + E_\mathbf{x}[\log(1 - \tilde{D}^{gan}(\mathbf{x}^p))]$ and $V_{\tilde{D}^{id}} = E_\mathbf{x}[\log \tilde{D}^{id}_{l_x^{id}}(\mathbf{x})]$, where $\tilde{D}_i^{id}$ is the i-th element in $\tilde{D}^{id}$.

For G, it has **two** training objectives as follows: **1)** Fool $\tilde{D}^{gan}$ to classify both of $\widehat{\mathbf{y}}$ and $\mathbf{x}^p$ as the real Domain A prototypes. **2)** Enable $\tilde{D}^{id}$ to predict the identity label of $\widehat{\mathbf{y}}$ as that of $\mathbf{y}$ (i.e., l_y^{id}), and the identity label of $\mathbf{x}^p$ as that of $\mathbf{x}$ (i.e., l_x^{id}). In light of the above two objectives, the ultimate objective function $\tilde{V}_G$ to train the generator G is formulated as

$$\max_G \tilde{V}_G = \tilde{V}_G^{gan} + \lambda_2 \tilde{V}_G^{id}, \tag{4}$$

where λ_2 is a balance parameter. $\tilde{V}_G^{gan}$ and $\tilde{V}_G^{id}$ are defined as $\tilde{V}_G^{gan} = E_{\mathbf{x},\mathbf{y}}[\log \tilde{D}^{gan}(\widehat{\mathbf{y}}) + \log \tilde{D}^{gan}(\mathbf{x}^p)]$ and $\tilde{V}_G^{id} = E_{\mathbf{x},\mathbf{y}}[\log \tilde{D}^{id}_{l_y^{id}}(\widehat{\mathbf{y}}) + \log \tilde{D}^{id}_{l_x^{id}}(\mathbf{x}^p)]$, respectively.

3.4 DisHPL Applications

After training, we can employ the trained generator G to handle the **three** applications. **1) Heterogeneous prototype learning:** Learning the Domain A (or Domain B) prototype from a Domain B (or Domain A) input image across domains. **2) Homogeneous prototype learning:** Learning the Domain A (or Domain B) prototype from a Domain A (or Domain B) input image within the same domain. **3) Heterogeneous face recognition:** Given a Domain

A (or Domain B) enrolment set and a new Domain B (or Domain A) query image, we can acquire their domain-invariant prototype features in the latent spaces and then perform classification.

4 EXPERIMENTAL RESULTS

4.1 Implementation Details

For G_{encA}, we use Lightened CNN [33] as the backbone for prototype feature extraction. For G_{encB}, we adopt a different deep neural network, i.e., CASIA-Net [35], as the backbone for extracting the domain features. G_{encA} encodes a 256-dimensional prototype feature while G_{encB} encodes a 50-dimensional domain feature. For G_{dec}, it takes a 306-dimensional feature vector as the input, and outputs a face image of 128×128 pixels. For D and $\tilde{D}$, they have the same network structure whose input is a face image of 128×128 pixels while the output is a $(N_d + 1)$-dimensional feature vector.

We optimize DisHPL by using the stochastic gradient descent with a mini-batch size of 5. We follow the work in [30] by adopting the Adam [14, 25] as the optimizer, in which the learning rate and momentum are set to be 0.0002 and 0.5, respectively. In the experiments, the four balance hyper-parameters, i.e., α_1 in Eqn. (1), λ_1 in Eqn. (2), α_2 in Eqn. (3), and λ_2 in Eqn. (4), are tuned via the grid search strategy and are all empirically set at 2.

4.2 Evaluation on Prototype Learning

This subsection evaluates the learnt heterogeneous and homogeneous prototypes by our DisHPL on BUAA NIR-VIS [8], CASIA NIR-VIS v2.0 [15], and CUFSF [36] datasets. On the two NIR-VIS datasets, we denote the NIR domain as Domain A, and the VIS domain as Domain B. On CUFSF, we denote the sketch domain as Domain A, and the photo domain as Domain B.

Given a random query image from Domain A, i.e., $\mathbf{x}$, and a random query image from Domain B, i.e., $\mathbf{y}$, DisHPL can generate four different prototypes: 1) the Domain A prototype of $\mathbf{x}$, i.e, $\mathbf{x}^p$, 2) the Domain B prototype of $\mathbf{x}$, i.e., $\widehat{\mathbf{x}}$, 3) the Domain A prototype of $\mathbf{y}$, i.e., $\widehat{\mathbf{y}}$, and 4) the Domain B prototype of $\mathbf{y}$, i.e., $\mathbf{y}^p$. In Figure 3, we illustrate five prototype learning examples of DisHPL on the above three datasets. It can be observed that, **1)** DisHPL successfully learns the *variation-free* heterogeneous prototypes across the NIR-to-VIS, VIS-to-NIR, sketch-to-photo and photo-to-sketch domains, as well as the homogeneous prototypes within the same VIS, NIR, sketch and photo domains. Intuitively, for these contaminated input images containing facial variations of different postures, expressions (e.g., surprise and happiness), and occlusion of glasses, DisHPL is capable of simultaneously transferring the domain styles and decreasing the facial variations; **2)** On the three datasets, most of the learnt homogeneous and heterogeneous prototypes by DisHPL capture well the identity characteristics of the contaminated input images, and look similar to the reference groundtruth prototypes.

4.3 Evaluation on HFR

This subsection evaluates DisHPL for HFR on the two NIR-VIS datasets. On BUAA NIR-VIS (or CASIA NIR-VIS v2.0), we choose 50 (or 360) identities for training while another 100 (or 358) identities for testing according to the standard evaluation protocol [3]. On each dataset,we choose one VIS image from each testing identity to build the enrolment set while all testing NIR images for querying.

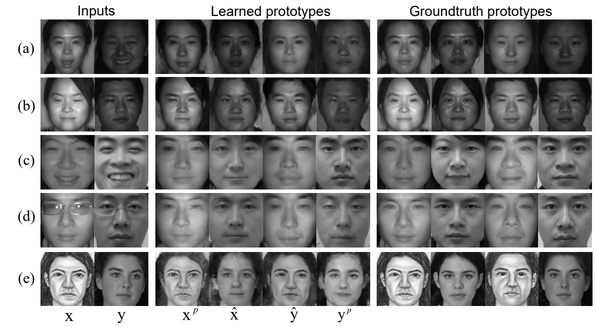

Figure 3: Prototype learning examples of DisHPL on BUAA NIR-VIS, CASIA NIR-VIS v2.0, and CUFSF datasets. In each row, figures from left to right are: the input query images x and y from two different domains, the four learnt prototypes by DisHPL, i.e., $\mathbf{x}^p$, $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$, and $\mathbf{y}^p$, and the corresponding groundtruth (GT) prototypes for reference.

Table 1: Recognition rates (%) of DisHPL and the compared feature learning-based methods on two NIR-VIS datasets.

Dataset	Hand-crafted			Deep Learning				Our
	KDSR	H2-LBP3	CEFD	TRIVET	ADFL	RGM	CAJL	DisHPL
BUAA	83.0	88.8	–	93.9	95.2	97.6	98.3	98.7
CASIA	37.5	43.8	85.6	95.7	98.2	97.2	–	97.3

In the HFR experiment, we select **7** representative NIR-VIS feature learning-based comparing approaches, involving **3** *handcrafted feature learning-based* KDSR [10], H2-LBP3 [24] and CEFD [6], and **4** *deep learning-based* ADFL [28], TRIVET [18], RGM [3] and CAJL [34]. The rank-1 recognition rates of DisHPL and the other methods on the two NIR-VIS datasets are listed in Table 1. We observe that, although DisHPL is not specifically designed for HFR, it still achieves promising performance, which indicates the learnt prototype features by DisHPL capture well the identity information across heterogeneous domains. Specifically, DisHPL performs the best on BUAA NIR-VIS dataset, and obtains comparable results to that of ADFL and RGM on CASIA NIR-VIS v2.0 dataset. The superiority of DisHPL owes to its two advantages: 1) the Max-Feature-Map based Lightened CNN in G_{encA} is adaptive to different appearances in different modalities [33]; and 2) the identity-relevant sub-discriminators D^{id} and $\tilde{D}^{id}$ force the encoders to accurately encode the identity information in the learnt prototype features.

5 CONCLUSION

This paper has studied an emerging challenging HPL problem, which involves two coupled subproblems of domain transfer and prototype learning. To tackle HPL, we have proposed the DisHPL to jointly address the above two subproblems in a unified disentanglement learning framework. Given a contaminated face image from the source domain, DisHPL is able to simultaneously: 1) disentangle its domain and prototype features, and 2) generate proper heterogeneous and homogeneous prototypes. Empirically studies on various NIR-VIS and sketch-photo face datasets have validated the effectiveness of DisHPL in both HPL and HFR tasks.

6 ACKNOWLEDGMENTS

This work was supported in part by NSFC under Grant: 61862043, NSFC/RGC Joint Research Scheme under Grant: N_HKBU214/21, and General Research Fund of RGC under Grant: 12201321.

REFERENCES

[1] 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 *CVPR*, pages 156–163, 2009.

[2] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang. Occlusion-aware face inpainting via generative adversarial networks. In *ICIP*, pages 1202–1206, 2017.

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

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

[5] X. Gao, N. Wang, D. Tao, and X. Li. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8):1213–1226, 2012.

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

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[8] D. Huang and e. a. Sun. The BUAA-VisNir face database instructions. *Technical Report IRIP-TR-12-FR-001, Beihang University*, 6, 2012.

[9] 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 *ICCV*, pages 2439–2448, 2017.

[10] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Transactions on Image Processing*, 22(1):353–362, 2013.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[12] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang. Disentangled representation learning for 3d face shape. In *CVPR*, pages 11957–11966, 2019.

[13] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee. Complete face recovery GAN: Unsupervised joint face rotation and de-occlusion from a single-view image. In *WACV*, pages 3711–3721, 2022.

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

[15] S. Li, D. Yi, Z. Lei, and S. Liao. The CASIA nir-vis 2.0 face database. In *CVPRW*, pages 348–353, 2013.

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

[17] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *CVPR*, volume 1, pages 1005–1010, 2005.

[18] X. Liu, L. Song, X. Wu, and T. Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *ICB*, pages 1–8, 2016.

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

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

[21] 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 Transactions on Information Forensics and Security*, 17:1544–1557, 2022.

[22] 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 Transactions on Neural Networks and Learning Systems*, 2021.

[23] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[24] M. Shao and Y. Fu. Cross-modality feature learning through generic hierarchical hyperlingual-words. *IEEE Transactions on Neural Networks and Learning Systems*, 28(2):451–463, 2016.

[25] R. Shao, X. Lan, J. Li, and P. C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031.

[26] R. Shao, P. Perera, P. C. Yuen, and V. M. Patel. Federated generalized face presentation attack detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[27] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan. Geometry guided adversarial facial expression synthesis. In *ACM MM*, pages 627–635, 2018.

[28] L. Song, M. Zhang, X. Wu, and R. He. Adversarial discriminative heterogeneous face recognition. In *AAAI*, volume 32, 2018.

[29] Y. Song, L. Bao, Q. Yang, and M.-H. Yang. Real-time exemplar-based face sketch synthesis. In *ECCV*, pages 800–813, 2014.

[30] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1415–1424, 2017.

[31] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3007–3021, 2018.

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

[33] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[34] M. Ye, W. Ruan, B. Du, and M. Z. Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021.

[35] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[36] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.

[37] H. Zhou, Z. Kuang, and K.-Y. K. Wong. Markov weight fields for face sketch synthesis. In *CVPR*, pages 1091–1097, 2012.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.