VD-GAN: A Unified Framework for Joint Prototype and Representation Learning From Contaminated Single Sample per Person

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

***Abstract*—Single sample per person (SSPP) face recognition with a *c*ontaminated biometric *e*nrolment database (SSPP-ce FR) is an emerging practical FR problem, where the SSPP in the enrolment database is no longer standard but contaminated by nuisance facial variations such as expression, lighting, pose, and disguise. In this case, the conventional SSPP FR methods, including the patch-based and generic learning methods, will suffer from serious performance degradation. Few recent methods were proposed to tackle SSPP-ce FR by *either* performing prototype learning on the contaminated enrolment database *or* learning discriminative representations that are robust against variation. Despite that, most of these approaches can only handle a specified *single* variation, e.g., pose, but cannot be extended to *multiple* variations. To address these two limitations, we propose a novel Variation Disentangling Generative Adversarial Network (VD-GAN) to *jointly* perform prototype learning and representation learning in a unified framework. The proposed VD-GAN consists of an encoder-decoder structural generator and a multi-task discriminator to handle universal variations including single, multiple, and even mixed variations in practice. The generator and discriminator play an adversarial game such that the generator learns a discriminative identity representation and generates an identity-preserved prototype for each face image, while the discriminator aims to predict face identity label, distinguish real vs. fake prototype, and disentangle target variations from the learned representations. Qualitative and quantitative evaluations on various real-world face datasets containing single/multiple and mixed variations demonstrate the effectiveness of VD-GAN.**

***Index Terms*—Single sample per person, prototype learning, representation learning, generative adversarial network.**

Manuscript received August 6, 2020; revised November 17, 2020 and December 23, 2020; accepted December 28, 2020. Date of publication January 8, 2021; date of current version February 9, 2021. The work of Yiu-ming Cheung was supported in part by the NSFC under Grant 61672444, in part by the Hong Kong Baptist University (HKBU) under Grant RC-FNRA-IG/18-19/SCI/03 and Grant RC-IRCMs/18-19/SCI/01, and in part by the Innovation and Technology Fund (ITF) of the Innovation and Technology Commission (ITC) of the Government of the Hong Kong, SAR, under Project ITS/339/18. The work of Bihan Wen was supported in part by the Ministry of Education, Republic of Singapore, under the Start-Up Grant; and in part by the National Research Foundation (NRF), Singapore, through the Singapore Cybersecurity Consortium (SGCSC) Grant Office, under Grant SGCSC_Grant_2019-S01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding author: Bihan Wen.)*

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

Binghui Wang and Yiran Chen are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: binghui.wang@duke.edu; yiran.chen@duke.edu).

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

Digital Object Identifier 10.1109/TIFS.2021.3050055

I. Introduction

SINGLE sample per person face recognition (SSPP FR), i.e., recognizing an identity based on his/her *single* image sample from the biometric enrolment database,[1] has attracted considerable attentions in information security owing to its potential applications in criminal identification, access control, video surveillance, person re-identification, to name a few [1]–[12]. Fisher-based methods [13]–[17] are inapplicable to SSPP FR as they require within-class information that is unavailable. Besides, the classic sparse representation-based classifier (SRC) [18] and collaborative representation-based classifier (CRC) [19] are also limited, due to insufficient enrolment data.

In the literature, there are two types of SSPP FR methods [20], namely the *patch-based* and *generic learning* methods. The patch-based methods [21]–[25] divide each single sample into several local patches for discriminative learning or recognition. In contrast, the generic learning methods [26]–[30] assume that a query sample is composed by the prototype and its intra-personal variations (i.e., the P+V model). The prototype is approximated by the original enrolment sample, while the variation dictionary is generated from an auxiliary generic set, which contains identities *not of interest* and encodes the difference between the query and enrolment samples.

However, all the above-mentioned SSPP FR methods assume that the biometric database contains only the standard enrolment samples with frontal pose and neutral expression, under normal lighting, and without occlusion/disguise (denoted as SSPP-se FR for short). In practice, many enrolment samples can be collected under less constrained environments, and are not standard faces anymore [9], [31]. For example, in criminal identification, various nuisance variations including expression, lighting, disguise, pose, and misalignment could exist in the enrolment samples of suspects such as the smugglers and illegal immigrants. Instead of the standard data acquisition at the police station, these enrolments are more likely provided by witnesses with unprocessed mobile photos or intercepted from low-quality surveillance videos. The SSPP FR problem with such a contaminated biometric

[1] More standardized biometric vocabularies can refer to the website of https://www.christoph-busch.de/standards.html

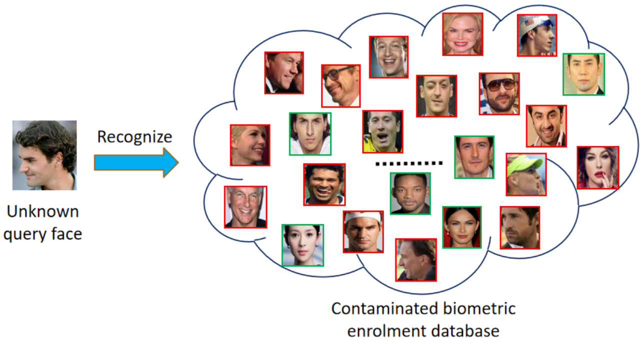


Fig. 1. Illustration of SSPP FR with a contaminated biometric enrolment database (i.e., SSPP-ce FR). The samples in the green box are standard enrolment samples while those in the red box are contaminated ones containing different variations. Better view in color version.

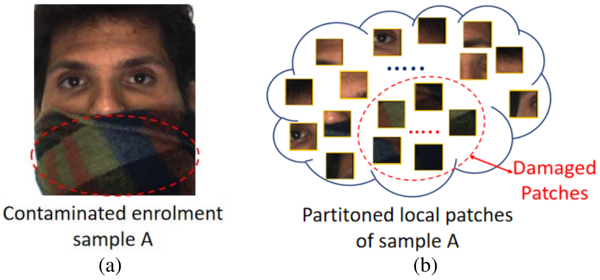


Fig. 2. (a) One contaminated enrolment sample wearing a scarf, and (b) its partitioned local patches. In these patches, some are damaged and capture useless information for discriminative learning or recognition.

enrolment database is termed as the SSPP-ce FR, as illustrated in Fig. 1.

There are two major challenges of the SSPP-ce FR problem: (i) a contaminated enrolment sample potentially yields an inaccurate prototype of the identity, and (ii) a query sample may be contaminated differently comparing to the enrolment of the same identity, which leads to an enlarged dissimilarity between them. The conventional SSPP FR methods can hardly tackle these challenges. For example, Fig. 2 illustrates that the patch-based methods are highly sensitive to local patch damages caused by sample contamination. Fig. 3 shows a failure case in which the generic learning methods mis-pair the query with the incorrect enrolment sample.

Most recent works attempt to tackle the SSPP-ce FR challenges, by either performing prototype learning on the contaminated enrolment database [31]–[38] or learning discriminative representations against variations for recognition [39]–[44]. While the prototype learning can generate realistic-looking prototypes for forensic experts, the discriminative representation learning can produce machine-readable features for FR. These two approaches are closely related: on the one hand, a highly discriminative representation can guide the generation of proper prototypes that capture the identity characteristics; on the other hand, an identity-preserved prototype can in turn strengthen the discriminant capability of the representation. Unfortunately, the two approaches have yet to be unified via an end-to-end scheme. Besides, most of the existing prototype learning or representation learning methods only consider a *single variation* such as pose or lighting, which are hard to generalize to multiple variations. In the literature, there are few multi-variation methods [31], [32], [39], [41], but are unsupervised for representation learning, or require the query set to be known in advance for prototype learning, which would be impractical for SSPP-ce FR.

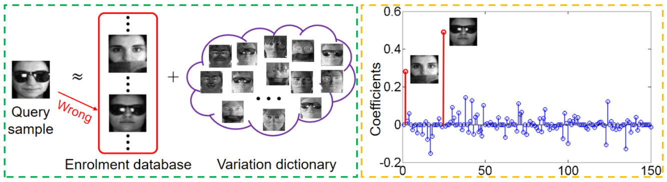


Fig. 3. A failed classification example of a typical P+V model-based generic learning method, i.e., SLRC [28], for SSPP-ce FR. In the example, a query sample wearing sunglasses is misclassified as an identity wearing the similar type of sunglasses (left), according to the representation coefficients (right).

To this end, we propose a novel Variation Disentangling scheme using Generative Adversarial Network (VD-GAN), to simultaneously learn prototypes and representations with a unified framework. Fig. 4 illustrates the architecture of the proposed VD-GAN, which consists of an encoder-decoder structural generator G and a multi-task discriminator $D = [D^{id}, D^{var}, D^{gan}]$ that predicts face identity, detects the existence (or not) of variation, and distinguishes real vs. fake prototype. G and D play an adversarial game: G strives for generating an identity-preserved prototype to fool D, while D guides G to preserve the identity information in the learned representation. Compared to the existing variation discriminators [42]–[44] which classify the style for a single variation (e.g., pose angles), our D^{var} is a binary classifier that is used to decide whether there exist variations or not and is applicable to universal variations. Thus, the learned representation by VD-GAN is discriminative in terms of identity and is disentangled w.r.t. nuisance variations, which is very suitable for the SSPP-ce FR task. Furthermore, VD-GAN introduces a unique reconstruction penalty in G to preserve the prototype generation from uncontaminated standard inputs, which can be beneficial to SSPP-ce FR. Extensive experiments are conducted to evaluate the effectiveness of VD-GAN over seven real-world face datasets (i.e., Multi-PIE, EYaleB, CAS-PEAL, FERET, AR, CFP, and LFW), which contain single, multiple, and mixed variations. Both quantitative and qualitative results show that VD-GAN can learn realistic-looking prototypes as well as discriminative identity representations that are disentangled from nuisance variations. Moreover, VD-GAN outperforms the state-of-the-art SSPP-ce FR methods over all evaluated datasets with promising improvements.

To the best of our knowledge, the proposed VD-GAN is the first work to jointly perform prototype learning and representation learning using a unified and end-to-end framework. Furthermore, compared to many existing works that can only handle a single variation in the samples, VD-GAN demonstrates robust performance against universal variations. The contributions of the paper are summarized as follows:

- We propose a novel VD-GAN, an end-to-end model for joint prototype and representation learning. VD-GAN can reconstruct the realistic-looking prototypes with samples from the contaminated biometric enrolment database.

- We propose a multi-task discriminator to assist in learning representations that are discriminative in terms of identity and are disentangled w.r.t. nuisance facial variations.
- We conduct extensive experiments on various real-world face datasets with single/multiple and mixed variations to demonstrate the effectiveness of VD-GAN for prototype and representation learning, as well as the superiority for SSPP-ce FR over the state-of-the-art counterparts.

The reminder of this article is organized as follows. Section II makes an overview of the related works, and Section III briefly reviews the vanilla GAN. Section IV details the proposed VD-GAN. In Section V, we perform experiments on seven real-world face datasets to evaluate the performance of VD-GAN. Finally, we draw a conclusion in Section VI.

II. Related Works

A. Conventional SSPP FR Methods

Conventional SSPP FR methods are designed for the classic SSPP-se FR problem and can be classified into the patch-based methods and generic learning methods.

For the patch-based methods, one can either integrate the classification outputs from the partitioned patches in a query sample for recognition, or perform discriminative learning on the patches in enrolment samples followed by the matching between the enrolment and query samples. For example, Wright *et al.* [18] and Zhu *et al.* [21] extended SRC and CRC to their patch-based versions, i.e., PSRC and PCRC, respectively, by combing the SRC or CRC outputs from the partitioned query patches. Lu *et al.* [22] presented a discriminative multi-manifold analysis (DMMA) method provided that patches of each enrolment identity lie in an individual manifold, thus converting FR to a manifold-to-manifold matching problem. However, these patch-based methods cannot generate auxiliary information and the discriminative learning from patches is highly sensitive to image variations [29]. Particularly in SSPP-ce FR, some damaged patches may even capture useless information for discriminative learning or recognition.

The generic learning methods usually leverage the popular P+V model [27] for recognition. Formally, given a query face sample $\mathbf{y}$, it can be represented as

$$\mathbf{y} = \mathbf{P}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where $\mathbf{P}$, $\mathbf{V}$, and $\mathbf{e}$ are the enrolment sample dictionary, the variation dictionary and a small noise, respectively, $\boldsymbol{\alpha}$ is the sparse coefficient vector choosing a few of enrolment samples (i.e., identities) from $\mathbf{P}$, and $\boldsymbol{\beta}$ is another sparse coefficient vector that selects a small subset of dictionary $\mathbf{V}$. Subsequently, the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be calculated through the following optimization problem:

$$\begin{bmatrix}\boldsymbol{\alpha}^*\\ \boldsymbol{\beta}^*\end{bmatrix} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\arg\min} \left\| \mathbf{y} - [\mathbf{P} \quad \mathbf{V}] \begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} \right\|_1, \tag{2}$$

where $||.||_2$ and $||.||_1$ indicate the l_2-norm and l_1-norm, respectively, and λ is a regularization parameter. Finally, similar to SRC, $\mathbf{y}$ will be classified into the enrolment sample (i.e., identity) with the smallest reconstruction residual. By virtue of the P+V model, Deng *et al.* [28] proposed a superposed linear representation classifier (SLRC) and generated the variation dictionary by subtracting average face from the samples of each identity in the generic set. Yang *et al.* [29] proposed a sparse variation dictionary learning (SVDL) method to use the relationship between the enrolment and generic samples. Ji *et al.* [30] extended SVDL by additionally using the contributions of different generic identities. However, these generic learning methods are not optimal for the new SSPP-ce FR problem because a contaminated enrolment sample can hardly be treated as an appropriate prototype for the P+V model.

B. Prototype Learning-Based Methods

Prototype learning-based methods aim to generate identity-preserved prototypes for contaminated enrolments, where the conventional SSPP FR methods can be applicable. The existing prototype learning-based methods are roughly classified into two types: the former is to exploit auxiliary information in query set for restoring contaminated sample in the enrolment database, while the latter is to train mappings between contaminated and standard samples in the generic set and then transfer them to the enrolment database for prototype learning.

For the former type, the typical methods are semi-supervised sparse representation based classification (S^{3}RC) [31] and iterative dynamic generic learning (IDGL) [32]. The two methods introduce the query set into the enrolment database, and then estimate the prototypes by the clustering centroid from a Gaussian mixture model (GMM) [45] or a semi-supervised low-rank representation [46]. Despite promising prototypes obtained by them, they require the join of unknown query set, which may be impractical from a real-time perspective.

For the latter type, benefiting from the powerful mapping ability of GAN, a series of GAN variants [33]–[38] have been proposed to decrease specified variations and to synthesize the corresponding identity-preserved prototypes. For example, Ma *et al.* [33] proposed a style translation GAN to learn the mappings between arbitrary lighting domains and standard lighting domain for normalization; Chen *et al.* [34] developed an occlusion-aware GAN to detect and recover missing regions in occluded or disguised samples; Song *et al.* [35] presented a geometry-guided GAN by using fiducial points to guide facial expression transfer and neutralization; Huang *et al.* [36] presented a two-pathway GAN to correct the ill-posed samples through both global and local transformations. Although these GAN variants perform well for the specified single variation such as lighting, disguise, expression or pose, they need to know the input pattern of the variation in advance and cannot handle unspecified multiple variations, which is unsuitable for SSPP-ce FR from the practical point of view.

C. Representation Learning-Based Methods

Representation learning is a major and hot topic in artificial intelligence, but how to design a reasonable objective for learning good representations is still an open-ended question [47].

Among a number of representation learning-based methods, some can be applied to SSPP-ce FR. In addition to traditional unsupervised subspace learning methods [48]–[51] such as principal component analysis (PCA) [48], increasing attentions have been given to the auto-encoder based methods [39]–[44], [52]. Vincent *et al.* [39] proposed a de-noising auto-encoder (DAE) to learn features which are robust to some predefined noises. Gao *et al.* [40] extended DAE to a stacked supervised auto-encoder to deal with realistic facial variations. Kingma and Welling [41] developed a variational auto-encoder (VAE) architecture to disentangle the factors of variation. Based on VAE, Kulkarni *et al.* [42] proposed a deep convolution inverse graphics network to generate representations disentangled w.r.t. pose and lighting. Furthermore, Tran *et al.* [43], [44] introduced GAN into auto-encoder and proposed a disentangled representation learning GAN (DR-GAN), which learns a pose-invariant representation and meanwhile rotating input face to a specified pose. The architecture of our VD-GAN is inspired by that of DR-GAN, but the purpose is different. In VD-GAN, we aim to learn neutral prototypes for contaminated samples but not performing face rotation. Besides, VD-GAN is designed for disentangling universal variations in the learned representations, but not limited to the pose variation.

III. Background on GAN

Goodfellow *et al.* [53] proposed the generative adversarial network (GAN) consisting of two main components, i.e., a generator G and a discriminator D, which play a minimax two-player game. The discriminator D is trained to distinguish between the real image $\mathbf{x}$ and the generated fake image $\hat{\mathbf{x}}$, while the generator G is trained to generate realistic-looking images, i.e., $G(\mathbf{z})$, from a random noise vector $\mathbf{z}$ to fool D. The optimization problem with the objective function of GAN is formulated as

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1-D(G(\mathbf{z})))], \quad (3)$$

where p_{data} and p_z denote the distributions of the training data and the noise $\mathbf{z}$, respectively. Note that the minimization of $\log(1-D(G(\mathbf{z})))$ can be replaced by the maximization of $\log(D(G(\mathbf{z})))$ to provide much stronger gradients early in learning [53]. Therefore, the objective in Eq. (3) can be reformulated as follows:

$$\max_D V_D(G,D) = E_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z}[\log(1-D(G(\mathbf{z})))], \quad (4)$$

$$\max_G V_G(G,D) = E_{\mathbf{z}\sim p_z}[\log(D(G(\mathbf{z})))]. \quad (5)$$

The generator G in Eq. (5) and discriminator D in Eq. (4) are iteratively updated until convergence is achieved or a predefined maximum number of iterations is reached.

IV. The Proposed Method

In this section, we start by defining the problem we are solving and the proposed objectives, followed by introducing the proposed VD-GAN with the network architecture, training and evaluation schemes.

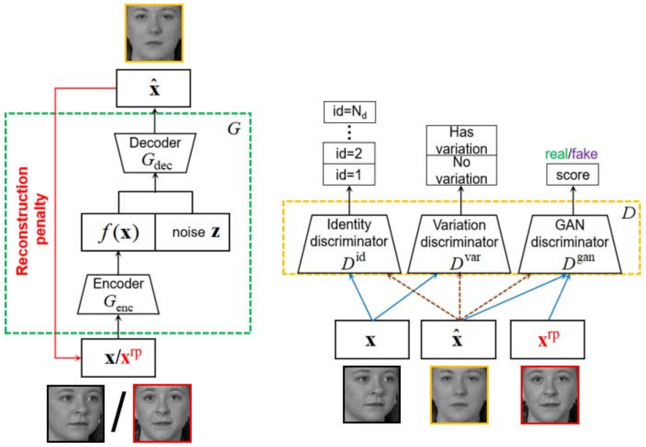


Fig. 4. The architecture of the proposed VD-GAN. $\mathbf{x}$, $\mathbf{x}^{rp}$, $\widehat{\mathbf{x}}$, and $f(\mathbf{x})$ denote the input face image, the standard/real prototype image, the generated prototype image, and the learned representation of $\mathbf{x}$, respectively. A unique reconstruction penalty is introduced in G to preserve the prototype generation from uncontaminated standard inputs. When training D, D^{id} and D^{var} predict the ID and variation of $\mathbf{x}$, respectively; D^{gan} aims to assign a high score to $\mathbf{x}^{rp}$ but a low score to $\widehat{\mathbf{x}}$ so as to distinguish real vs. fake prototype. When training G, $\widehat{\mathbf{x}}$ aims to fool D^{id}, D^{var}, and D^{gan} to classify it into the ID of $\mathbf{x}$, to judge it with no variation, and to assign it a high score of being real prototype, respectively.

A. Problem and Objectives

We propose to jointly conduct the prototype and representation learning for face images using a unified framework. To be specific, for an input face image $\mathbf{x}$, the proposed VG-GAN generates a prototype $\widehat{\mathbf{x}}$ and a discriminative identity representation $f(\mathbf{x})$ to achieve the following objectives:

- **Prototype learning**: to reconstruct a high-quality (i.e., realistic-looking) prototype $\widehat{\mathbf{x}}$ for the input face image $\mathbf{x}$, such that $\widehat{\mathbf{x}}$ 1) is variation-free, and 2) preserves the individual characteristics of $\mathbf{x}$.
- **Representation learning**: to learn a discriminative identity representation $f(\mathbf{x})$ for $\mathbf{x}$, such that $f(\mathbf{x})$ 1) captures the identity information of $\mathbf{x}$, and 2) is invariant to any facial variations in $\mathbf{x}$.

It is clear that both the prototype learning and representation learning need to be robust to variations and preserve face identity. The two learning tasks are optimized by an end-to-end training from data with the annotations $\mathbf{y} = \{y^{id}, y^{var}\}$, where y^{id} and y^{var} denote the face identity and whether the face contains variation (i.e., y^{var} is a binary label), respectively. Besides, the proposed multi-task discriminator ensures $\widehat{\mathbf{x}}$ to be realistic-looking via adversarial learning.

B. VD-GAN

We propose a variation disentangling scheme using GAN (VD-GAN), whose architecture is illustrated in Fig. 4. The proposed VD-GAN consists of two main parts: an encoder-decoder network serving as the generator G and a multi-task discriminator D. In the following, we first introduce the generator G and the discriminator D, and then present the alternative training scheme between G and D.

1) Generator G: The proposed generator G is composed of an encoder G_{enc} and a decoder G_{dec}. Given an input face image $\mathbf{x}$, G_{enc} aims to learn an identity representation $f(\mathbf{x}) = G_{enc}(\mathbf{x})$ for $\mathbf{x}$, while G_{dec} aims to synthesize the prototype image $\widehat{\mathbf{x}}$ such that it has the same identity as $\mathbf{x}$. Specifically, G_{dec} takes the concatenation of the representation $f(\mathbf{x})$ and a random noise vector $\mathbf{z} \in R^{Nz}$ sampled from a distribution p_z as the input, then generates the prototype image $\widehat{\mathbf{x}} = G_{dec}(f(\mathbf{x}), \mathbf{z})$. Here, the noise vector $\mathbf{z}$, which we draw from a uniform distribution $[-1, 1]^{Nz}$, is to enhance the robustness and generalizability of the trained generator.

2) Discriminator D: The proposed multi-task discriminator D consists of three sub-discriminators, namely D^{id}, D^{var} and D^{gan}. To be specific, they are:

1) D^{id} outputs a N_d-dimensional vector for identity classification with N_d as the total number of identities in the training set.
2) D^{var} is a binary classifier to decide whether the target variation exists in an input image.
3) D^{gan} is a standard GAN discriminator to distinguish the real prototype vs. fake prototype generated by the generator G. More specifically, D^{gan} assigns a score to each image and a higher score indicates that the image is closer to the real prototype image.

3) VD-GAN Training: Suppose we are given a training set of N_d identities with each face image $\mathbf{x}$ annotated by the label $\mathbf{y} = \{y^{id}, y^{var}\}$, where y^{id} and y^{var} denote the face identity and whether the face contains variation, respectively. We also collect standard images (i.e., images not corrupted by variations) in the training set to form the real prototype corpus. We denote each standard/real prototype image as $\mathbf{x}^{rp}$, and its distribution as p_{real}. As a comparison, we denote that all face images $\mathbf{x}$ in the training set are sampled from the distribution p_{data}, i.e., $\mathbf{x} \sim p_{data}$.

For the generator G, we have the following four objectives:

- Enable D^{id} to classify the generated prototype $\hat{\mathbf{x}}$ by G as the same identity label as the input image $\mathbf{x}$, i.e., y^{id}.
- Enable D^{var} to detect that no variation exists in the generated prototype $\hat{\mathbf{x}}$.
- Fool D^{gan} to classify the generated *fake* prototype $\hat{\mathbf{x}}$ as a real prototype, i.e., G makes D^{gan} assign a high score to $\hat{\mathbf{x}}$ of being real prototype.
- Enable the generated prototype to well reconstruct the standard input image. That is, for each input image not corrupted by variations, G keeps the corresponding generated prototype be the same as this input.

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

$$\max_G V_G = V_G^{gan} + \mu_1 V_G^{id} + \mu_2 V_G^{var} - \mu_3 V_G^{rec}, \tag{6}$$

where μ_1, μ_2, and μ_3 are the weighting hyper-parameters for the hybrid objective V_G. The four sub-objectives V_G^{id}, V_G^{var}, V_G^{gan} and V_G^{rec} are defined as follows:

$$V_G^{id}(G, D^{id}, \mathbf{x}, \mathbf{z}) = E_{\mathbf{x},\mathbf{y},\mathbf{z}}[\log D^{id}_{y^{id}}(G(\mathbf{x}, \mathbf{z}))], \tag{7}$$

$$V_G^{var}(G, D^{var}, \mathbf{x}, \mathbf{z}) = E_{\mathbf{x},\mathbf{y},\mathbf{z}}[\log D^{var}_{y^{var}}(G(\mathbf{x}, \mathbf{z}))], \tag{8}$$

$$V_G^{gan}(G, D^{gan}, \mathbf{x}, \mathbf{z}) = E_{\mathbf{x},\mathbf{z}}[\log D^{gan}(G(\mathbf{x}, \mathbf{z}))], \tag{9}$$

$$V_G^{rec}(G, \mathbf{x}^{rp}, \mathbf{z}) = E_{\mathbf{x}^{rp},\mathbf{z}}[\frac{1}{2}||\mathbf{x}^{rp} - G(\mathbf{x}^{rp}, \mathbf{z})||_F^2], \tag{10}$$

where D_i^{id} and D_i^{var} denote the i-th element in D^{id} and D^{var}, respectively; $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}^{rp}$ and $\mathbf{z}$ are sampled from their respected distributions, i.e., $\mathbf{x}, \mathbf{y} \sim p_{data}$, $\mathbf{x}^{rp} \sim p_{real}$, $\mathbf{z} \sim p_z$, $\mathbf{y} = [y^{id}, y^{var}]$; and $||.||_F$ denotes the Frobenius norm. Note that the reconstruction loss in Eq. (10) is to enable the generated prototype image of $\mathbf{x}^{rp}$, i.e., $G(\mathbf{x}^{rp}, \mathbf{z})$, to be close to $\mathbf{x}^{rp}$ at the pixel level.

For the discriminator D, we have the following three objectives:

- Given the input image $\mathbf{x}$, D^{id} aims to correctly predict its identity label y^{id}.
- Given the input image $\mathbf{x}$, D^{var} aims to correctly predict its variation label y^{var}, which indicates the existence (or not) of any facial variation.
- Given the *real* prototype image $\mathbf{x}^{rp}$ and the generated *fake* prototype image by G, i.e., $\widehat{\mathbf{x}} = G(\mathbf{x}, \mathbf{z})$, D^{gan} aims to classify $\mathbf{x}^{rp}$ as the real prototype and classify $\hat{\mathbf{x}}$ as the fake prototype.

Formally, our final objective function V_D for training $D = [D^{id}, D^{var}, D^{gan}]$ is as follows:

$$\max_D V_D = V_D^{gan} + \lambda_1 V_D^{id} + \lambda_2 V_D^{var}, \tag{11}$$

where λ_1 and λ_2 are two trade-off parameters, and V_D^{id}, V_D^{var} and V_D^{gan} are defined as follows:

$$V_D^{id}(D^{id}, \mathbf{x}) = E_{\mathbf{x},\mathbf{y}}[\log D^{id}_{y^{id}}(\mathbf{x})], \tag{12}$$

$$V_D^{var}(D^{var}, \mathbf{x}) = E_{\mathbf{x},\mathbf{y}}[\log D^{var}_{y^{var}}(\mathbf{x})], \tag{13}$$

$$\begin{aligned} V_D^{gan}(G, D^{gan}, \mathbf{x}^{rp}, \mathbf{x}, \mathbf{z}) = {} & E_{\mathbf{x}^{rp}}[\log D^{gan}(\mathbf{x}^{rp})] \\ & + E_{\mathbf{x},\mathbf{z}}[\log(1 - D^{gan}(G(\mathbf{x}, \mathbf{z})))]. \end{aligned} \tag{14}$$

We alternatively train the generator G and the discriminator D by solving the objective functions V_G in Eq. (6) and V_D in Eq. (11) iteratively. During the alternative training process, the generator G and the discriminator D will improve each other. On the one hand, with D being more powerful in classifying identity labels, judging the existence of target variations, and distinguishing real vs. fake prototype images, G strives for generating an identity-preserved prototype image in order to fool D. On the other hand, the increasing powerful D^{id} can guide G_{enc} to learn a discriminative representation that encodes as much identity information as possible. Meanwhile, D^{gan} and D^{var} cooperatively disentangle the variations in the learned representation, thus forcing the generated prototype to contain as little variation as possible.

4) Applications: In testing, with the trained VD-GAN model, we can leverage our generator G to do the following tasks:

1) Generating realistic-looking prototypes (e.g., an ID photo) for contaminated enrolment samples in testing set.
2) Obtaining the discriminative identity representations for both enrolment samples and query samples in testing set, which are robust against nuisance facial variations.
3) Applying the discriminative identity representations to perform the challenging SSPP-ce FR.

We will demonstrate the effectiveness of the proposed VD-GAN regarding these potential applications in Section V.

V. Experiments

In this section, we first explain the detailed experimental settings, and then demonstrate the effectiveness of the proposed VD-GAN by presenting the experimental results:

1) In Subsection V-B, we perform experiments on five benchmark datasets (i.e., Multi-PIE, E-YaleB, CAS-PEAL, FERET and AR) to qualitatively and quantitatively evaluate the learned prototypes by VD-GAN with four major single variations, i.e., expression, lighting, disguise and pose, and multiple variations.
2) In Subsection V-C, we evaluate the learned representations by VD-GAN on the above five benchmark datasets for SSPP-ce FR.
3) In Subsection V-D, we perform ablation study to investigate the importances of the identity discriminator (D^{id}), GAN discriminator (D^{gan}), variation discriminator (D^{var}), and the reconstruction penalty in G (G^{rec}), respectively.
4) In Subsection V-E, we explore the feasibility of VD-GAN to handle mixed variations on the unconstrained Celebrities in Frontal-Profile (CFP) and Labeled Faces in the Wild (LFW) datasets.

A. Experimental Settings

1) Dataset Description: Multi-PIE [54] consists of 337 identities with each containing face images with 6 different expressions across four sessions (Session 1-4), 15 poses, and 20 illuminations. We use a subset of 141 identities only containing expression variations, where 100 identities are randomly chosen for training and the rest 41 ones for testing.

EYaleB [55] consists of 38 identities under various lighting conditions and is classified into five subsets. Subset 1, Subsets 2-3 and Subsets 4-5 depict normal, slight-to-moderate, and severe lighting variations, respectively. Furthermore, we introduce the AR lighting subset into EYaleB to enrich the lighting variations as well as to expand the number of identities. On this EYaleB&AR Light dataset, we randomly choose 100 identities for training and the rest 38 ones for testing.

CAS-PEAL [56] consists of 1,040 identities with variations including accessory, facing direction, age, etc., which is believed to be the largest public dataset with occluded face images available. We use a subset of 300 identities from the normal and accessory categories, thus each identity has 1 neutral image and 6 images wearing different glasses and hats. We randomly choose 200 identities for training and the rest 100 ones for testing.

TABLE I

The Network Structures of G_{enc}, G_{dec} and D

G_{enc} and D		
Layer	Filter / Stride / Pad	Output Size
Conv1	3 × 3 / 1 / 1	96×96×32
Conv2	3 × 3 / 1 / 1	96×96×64
Conv3	4 × 4 / 2 / 1	48×48×64
Conv4	3 × 3 / 1 / 1	48×48×128
Conv5	4 × 4 / 2 / 1	24×24×128
Conv6	3 × 3 / 1 / 1	24×24×256
Conv7	4 × 4 / 2 / 1	12×12×256
Conv8	4 × 4 / 2 / 1	6×6×256
Conv9	3 × 3 / 1 / 1	6×6×N_f
AvgPool	6 × 6 / 1 / 0	N_f×1×1
FC (D only)	–	N_d+3

G_{dec}		
Layer	Filter / Stride / Pad	Output Size
FC	–	6×6×320
DeConv1	3 × 3 / 1 / 1	6×6×256
DeConv2	4 × 4 / 2 / 1	12×12×256
DeConv3	4 × 4 / 2 / 1	24×24×256
DeConv4	3 × 3 / 1 / 1	24×24×128
DeConv5	4 × 4 / 2 / 1	48×48×128
DeConv6	3 × 3 / 1 / 1	48×48×64
DeConv7	4 × 4 / 2 / 1	96×96×64
DeConv8	3 × 3 / 1 / 1	96×96×32
DeConv9	3 × 3 / 1 / 1	96×96×3

FERET [57] consists of 1,199 identities across ethnicity, gender, and age. We leverage a subset of 200 identities from five categories (“ba”, “be”, “bd”, “bf”, and “bg”) only containing pose variations, where 150 identities are randomly chosen for training and the rest 50 ones for testing.

AR [58] consists of 126 identities from two sessions with each identity having 26 face images with different facial variations. We use a subset of 100 identities containing multiple variations including expressions, illuminations and disguises (wearing sunglasses and scarf). We randomly choose 50 identities for training and the rest 50 ones for testing.

CFP [59] consists of 500 identities, each with 14 in-the-wild images collected in unconstrained environments. The face images show complex mixed variations including the combinations of poses&expressions, disguises&expressions, poses&lightings, etc. We leverage a subset of 3500 images of 350 identities, where 100 identities containing neutral images are chosen for testing and the rest 200 ones for training.

LFW [60] consists of over 13,000 images of 5,749 identities collected under uncontrolled environments with large variations in expressions, poses, illuminations, etc. We use a subset of 158 identities with no less than 10 images per identity from LFW-a, which is the aligned version of LFW, for evaluation. We choose 50 identities containing neutral images for testing and use the rest 108 ones for training.

Fig. 5 shows some gray face samples on Multi-PIE, EYaleB, CAS-PEAL, FERET, AR, CFP, and LFW datasets.

2) Implementation Details: We first present the network structures of G (including G_{enc} and G_{dec}) and D of VD-GAN in Table I. For G_{enc} and G_{dec}, note that batch normalization (BN) and exponential linear unit (ELU) are used after each convolutional layer and de-convolutional layer. G_{enc} and G_{dec} are bridged by the to-be-learned identity representation

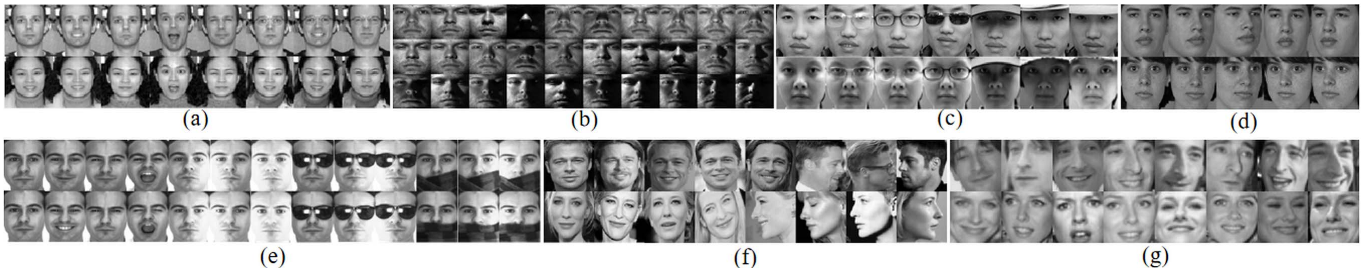


Fig. 5. Example gray face images from seven constrained and unconstrained datasets: (a) Multi-PIE; (b) EYaleB; (c) CAS-PEAL; (d) FERET; (e) AR; (f) CFP; (g) LFW.

TABLE II

DATASET PARTITION AND PARAMETER SETTING

Dataset	#Train. identity	#Test. identity	N_d	N_f, N_z	Trade-off parameter
Multi-PIE	100	41	100	N_f=320 N_z=50	λ_1=μ_1=2.0 λ_2=μ_2=0.5 μ_3=0.1
EYaleB&AR Light	100	38	100		
CAS-PEAL	200	100	200		
FERET	150	50	150		
AR	50	50	50		
CFP	200	100	200		

$f(\mathbf{x}) \in R^{N_f}$, which is the output of the AvgPool layer. Then, $f(\mathbf{x})$ is concatenated with a random noise $\mathbf{z} \in R^{N_z}$ and fed to G_{dec} to synthesize the prototype for $\mathbf{x}$. For the discriminator D, it has an extra fully connection (FC) layer compared to G_{enc}, and the output of the final FC layer is a (N_d+3)-dimensional vector. Specifically, the first N_d elements are the outputs of D^{id} and are used to predict the face identity label, the next two elements are for D^{var} to judge the existence of the target variation, and the last one element is reserved for D^{gan} to distinguish real vs. fake prototype images.

We train VD-GAN [2] by the mini-batch stochastic gradient descent (SGD) with a mini-batch size of 16. All weights are initialized from a zero-centered Normal distribution with the standard deviation of 0.02. We use the Adam optimizer [61] with tuned hyperparameters for optimizing, where the learning rate and momentum are empirically set at 0.0002 and 0.5, respectively, as suggested in [43] and [62].

3) Parameter Setting: For each evaluated dataset, N_d is set as the total number of identities in the training set, the dimensions of the learned representation N_f and the noise vector N_z are fixed at 320 and 50, respectively. We tune all trade-off hyper-parameters via grid search. Specifically, we observe that VD-GAN achieves promising performance when the trade-off parameters λ_1, λ_2 in Eq. (11), μ_1, μ_2, μ_3 in Eq. (6) are set at 2.0, 0.5, 2.0, 0.5, 0.1, respectively, and fix their values across all datasets. Moreover, the number of training and testing identities on each dataset are also specified.

All the above parameter settings and dataset partition are detailed in Table II.

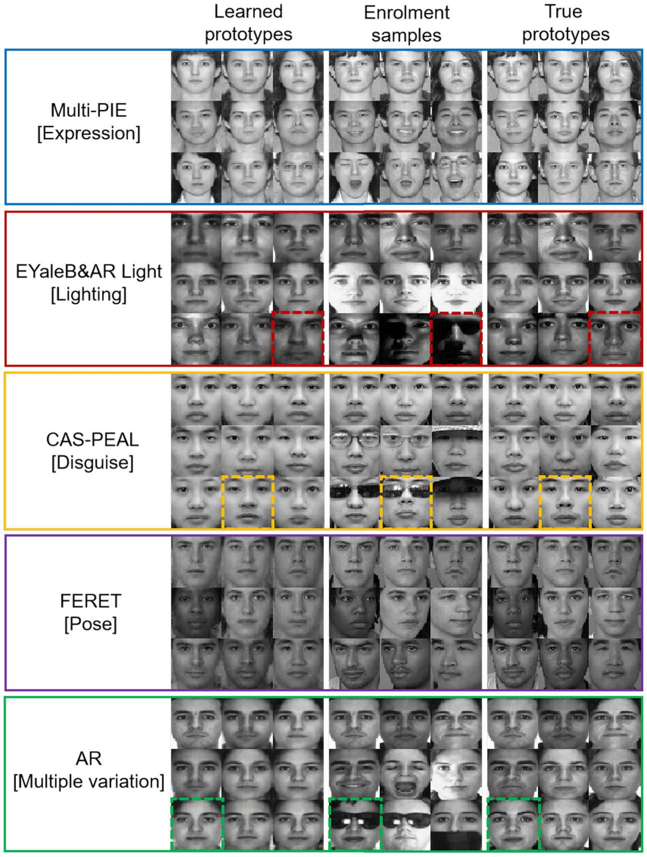


Fig. 6. Learned prototypes of some randomly selected examples on the Multi-PIE, EYaleB&AR Light, CAS-PEAL, FERET, and AR datasets. Figures from left to right are: learned prototypes by our VD-GAN, original enrolment samples, and true prototypes for reference.

B. Evaluation of the Learned Prototypes on Single/Multiple Variations

In this subsection, we first evaluate the learned prototypes by VD-GAN on Multi-PIE, E-YaleB&AR Light, CAS-PEAL, and FERET datasets, with each containing the single variation of expression, lighting, disguise or pose, respectively. Furthermore, we investigate the performance of VD-GAN on the AR dataset containing multiple variations. In the experiments, the quality of the learned prototypes is measured from both qualitative and quantitative perspective.

1) Qualitative Analysis Results: We show the learned prototypes for nine random enrolment samples on each dataset in Fig. 6. Among the nine selected samples, the first three are standard while the rest six are contaminated with different degrees of variations. For reference, we also provide the true prototypes of these enrolment samples. From Fig. 6, we have the following key observations:

[2] The code is released at https://github.com/PangMeng92/VD-GAN.git.

TABLE III
THE VERIFICATION PERFORMANCE OF VD-GAN ON THE MULTI-PIE, EYALEB&AR LIGHT, CAS-PEAL, FERET AND AR DATASETS

Dataset	AP (%)		TPR(%)@FAR=0.1	
	Baseline	VD-GAN	Baseline	VD-GAN
Multi-PIE [Expression]	77.4±1.1	**78.2±0.9**	60.8±1.0	**61.4±2.5**
EYaleB&AR Light [Lighting]	75.6±0.9	**88.2±1.3**	63.9±2.2	**78.7±3.4**
CAS-PEAL [Disguise]	66.3±1.9	**75.9±1.1**	49.7±3.2	**59.1±3.6**
FERET [Pose]	66.9±2.2	**75.3±0.8**	44.6±3.1	**57.5±1.1**
AR [Multiple variations]	60.2±2.6	**67.3±1.1**	37.0±3.4	**46.2±2.3**

1) For all standard enrolment samples on the five datasets, our VD-GAN enables to learn prototypes that are nearly the same as the true prototypes, owing to the reconstruction penalty.
2) For enrolment samples contaminated by a single variation such as expression, lighting, disguise or pose, VD-GAN successfully removes the corresponding variation in the learned prototypes. Moreover, even in the case that AR is contaminated by multiple variations and the input type of variation is unknown in advance, our VD-GAN is still able to learn almost variation-free prototypes.
3) In a few extreme cases where enrolment samples are contaminated by serious facial variations such as severe shadow or disguise of sunglasses, the generated prototypes seem inaccurate in terms of identity (see the examples surrounded by the dotted lines). This is because that some key information is missing in these severely damaged regions.

2) Quantitative Analysis Results: As most of the learned prototypes by our VD-GAN are visually appealing, it is expected that the learned prototypes are more appropriate to represent the identities than the original contaminated enrolment samples. To verify this assumption, we further perform quantitative analysis and conduct verification experiments on the learned prototypes by VD-GAN and true prototypes. Specifically, on each dataset, we randomly sample 600 pairs of learned prototypes and true prototypes, where 200 pairs are positive and the remaining 400 pairs are negative, for verification. For comparison, we treat the verification results between the original enrolment samples and the true prototypes as a baseline. The cosine similarity between each pair of samples is used for verification.

Two common metrics, i.e., average precision (AP) and true positive rate (TPR), are employed to measure the verification performance. For TPR, we tune the (cosine) similarity threshold such that the false acceptance rate (FAR) equals 0.1. Please refer to [63], [64] for the detailed definitions of the two metrics. Each verification experiment is repeated 5 times and the average results (± standard errors) are reported in Table III. It can be observed that, our VD-GAN consistently achieves better verification performance than the baseline method in all cases, especially on the E-YaleB&AR Light, CAS-PEAL and FERET face datasets where the differences between the true prototypes and contaminated samples are relatively large. For example, VD-GAN outperforms the baseline method by 12.6%, 9.6%, and 8.4% w.r.t. AP on E-YaleB&AR Light, CAS-PEAL, and FERET, respectively. The promising quantitative results validate that the learned prototypes by VD-GAN are closer to the true prototypes than the original contaminated enrolment samples.

C. Evaluation of the Learned Representations on Single/Multiple Variations

In this subsection, we evaluate the learned identity representations by VD-GAN for SSPP-ce FR on the Multi-PIE, E-YaleB&AR Light, CAS-PEAL, FERET and AR datasets.

On each dataset, we randomly choose one sample (could be a standard sample or a contaminated sample) for each identity from the testing set to construct the contaminated enrolment database, and use the rest ones as the query samples for recognition. We set the contaminated ratio (i.e., #contaminated samples / #total identities) ranging from 10% to 90% with an interval of 20%. Each experiment is repeated 5 times and the average result is reported. For comparison, we also report the results when the contaminated ratio is zero, which is exactly the setting of the SSPP-se FR problem.

As our testing scenarios include the specified single variation and unspecified multiple variations, we require that the evaluated methods should be able to handle universal variations. Hence, in our experiments, we choose 9 universal methods for comparison, including 2 representation learning-based methods, i.e., PCA [48] and VAE [41], 2 representation-based classifiers, i.e., SRC [18] and CRC [19], 2 well-known patch-based methods, i.e., PCRC [21] and DMMA [22], 2 recent generic learning methods, i.e., SVDL [29] and SLRC [28], and the state-of-the-art prototype learning-based S^3RC [31]. For SVDL, SLRC and S^3RC, the training set with multiple samples per identity is used as the generic set for generating variation dictionaries.

Regarding the parameter settings, the non-overlapped patch size for DMMA and PCRC is empirically set as 16×16 pixels. In addition, the other parameters of k_1, k_2, k, and σ in DMMA are set to be 30, 2, 2, and 100, respectively, according to the suggestion in [22]. The value of the regularization parameter λ of SRC, CRC, SLRC and S^3RC is fixed as 0.01. For SVDL, in accordance with [29], the parameters λ_1, λ_2 and λ_3 are set to be 0.001, 0.01 and 0.0001, respectively. For PCA, VAE and VD-GAN, the cosine distance metric is used for measuring the similarity between two learned representations, and the nearest neighbor classifier is adopted for classification.

Table IV presents the top-1 recognition accuracies of all the compared methods on the five datasets for SSPP-ce FR. We have the following observations:

1) VD-GAN consistently outperforms all the compared methods on the five datasets.
2) When there exists no contamination (ratio=0%) in the enrolment database, the two generic learning methods, i.e., SVDL and SLRC, obtain comparable results with VD-GAN on the Multi-PIE, EYaleB&AR Light, and CAS-PEAL datasets, but perform poorly on the FERET dataset. The reason is that the P+V model used in SVDL

and SLRC is a linear-based superposition model, which can handle linear expression, lighting, or disguise variation, but is less effective to deal with the non-linear pose variation. In contrast, our VD-GAN achieves promising performance on FERET because it is designed for disentangling universal variations (including pose) in the learned representations, as well as preserving the identity information by the identity discriminator. Note that with no contamination, S^{3}RC does not require the prototype recovery and degenerates to a P+V-based generic learning method. Therefore, it achieves similar performance with SVDL and SLRC.

3) When the enrolment database contains contaminated samples (ratio$>$0%), the accuracies of all methods have a tendency of decreasing as the contaminated ratio increases. However, in comparison with all the other methods, VD-GAN shows greater robustness to tolerate the contaminated ratio and the superiority is more obvious when the contaminated ratio is higher. For example, when the contamination ratio increases from 10% to 90%, VD-GAN has a gain over the second best method from 0.1% to 1.1% on Multi-PIE, from 2.5% to 13.4% on EYaleB&AR Light, from 2.2% to 6.4% on CAS-PEAL, from 15.7% to 25.2% on FERET, and from 0.6% to 5.9% on AR, respectively. The advantages of VD-GAN attribute to two key factors. First, for contaminated samples, the three discriminators, i.e., D^{id}, D^{var} and D^{gan}, work cooperatively to force the generator to encode as much identity information as possible (by D^{id}) but as little variation as possible (by D^{var} and D^{gan}) in the learned representation. Second, for standard samples, VD-GAN introduces a unique reconstruction penalty to further strengthen the learned representation.
4) S^{3}RC achieves higher recognition accuracies than the generic learning methods (i.e., SVDL and SLRC) with contamination. This is because S^{3}RC involves a prototype learning step for restoring contaminated enrolment samples. However, S^{3}RC performs poorly on EYaleB&AR Light face dataset. The reason is that, the quality of the learned prototypes by S^{3}RC relies heavily on the clustering performance of GMM, which is sensitive to severe lightings and shadows.
5) SVDL and SLRC obtain similar accuracies because they both use the P+V model for recognition. However, their accuracies are much lower than VD-GAN's and the gap is larger as the contaminated ratio increases. This observation validates that the classic P+V model is suboptimal for SSPP-ce FR when the enrolment samples are contaminated by variations.
6) The patch-based DMMA and PCRC perform better than the conventional SRC and CRC methods, but are inferior to the generic learning SVDL and SLRC methods. We find that DMMA or PCRC has its own advantages under different variations. For example, PCRC is more robust against lighting variation than DMMA, but is more sensitive to pose variation.

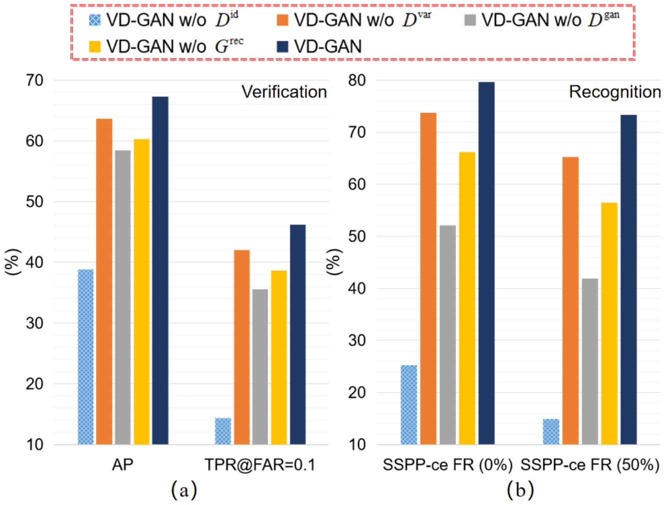


Fig. 7. Comparison results of VD-GAN and its four variants VD-GAN w/o D^{id}, VD-GAN w/o D^{var}, VD-GAN w/o D^{gan}, and VD-GAN w/o G^{rec} on the AR dataset. (a) Verification results. (b) Recognition results.

7) The representation learning-based VAE outperforms PCA as it also performs variation disentanglement during encoding. However, VAE is still not competitive with our VD-GAN because it is an unsupervised method, i.e., it does not exploit the labeled identity information.

In summary, the promising performance of VD-GAN demonstrates the effectiveness of its learned representations on single/multiple variations for SSPP-ce FR, especially when the contaminated ratio of the enrolment database is high.

D. Ablation Study on VD-GAN

In this subsection, we perform an ablation study on VD-GAN. In VD-GAN, D includes three components, i.e., D^{id}, D^{var}, and D^{gan}, for classifying identities, judging the existence of variation, and distinguishing real prototype vs. fake prototype, respectively. G includes a G^{rec} aiming to minimize the reconstruction loss for standard images. In this experiment, we aim to study the role of these components on VD-GAN's performance. Accordingly, we construct four variants of VD-GAN by removing D^{id}, D^{var}, D^{gan}, and G^{rec}, and denote them as VD-GAN w/o D^{id}, VD-GAN w/o D^{var}, VD-GAN w/o D^{gan}, and VD-GAN w/o G^{rec}, respectively.

Subsequently, we compare VD-GAN with the four variants in terms of: 1) the verification results of the learned prototypes; and 2) the recognition results of the learned representations on the AR dataset that contains multiple variations. As shown in Fig. 7, VD-GAN consistently outperforms all the four variants in both tasks. For example, VD-GAN has a gain over VD-GAN w/o D^{id}, VD-GAN w/o D^{var}, VD-GAN w/o D^{gan} and VD-GAN w/o G^{rec} by 28.5% (or 58.5%), 3.7% (or 8.1%), 8.9% (or 31.4%) and 7.0% (or 16.9%), respectively, w.r.t. AP (or recognition rate for SSPP-ce FR with the contaminated ratio of 50%). Our results verify that all the four components are necessary and help improve the performance of VD-GAN.

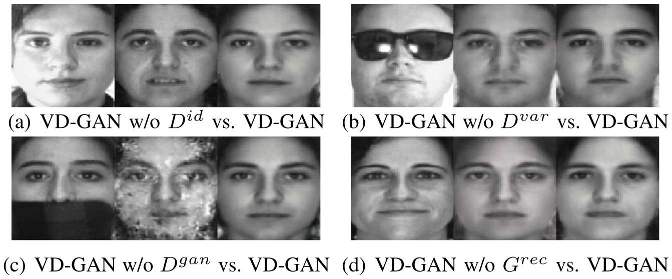
Moreover, we observe that different components have different impacts on the performance. Specifically, VD-GAN w/o D^{id} suffers the largest performance degradation, which

TABLE IV

RECOGNITION ACCURACIES (%) AND STANDARD ERRORS (%) OF DIFFERENT METHODS ON THE MULTI-PIE, E-YALEB&AR LIGHT, CAS-PEAL, FERET AND AR DATASETS FOR SSPP-CE FR. IN THE BRACKETS, WE SHOW THE IMPROVEMENT OF OUR VD-GAN OVER THE SECOND BEST METHOD

enrolment database		Representation learning		Representation-based classifier		Patch-based learning		Generic learning		Prototype learning	Our method
		PCA	VAE	SRC	CRC	DMMA	PCRC	SVDL	SLRC	S^3RC	VD-GAN
Multi-PIE [Expression]	0%	58.9	59.0	67.8	73.6	72.9	69.1	78.4	75.9	**79.8**	78.9
	10%	57.2±0.9	57.0±1.5	66.7±2.6	72.2±2.5	72.6±1.7	68.0±2.0	77.3±1.4	75.2±2.4	77.9±2.1	**78.0±1.3** (↑ 0.1)
	30%	56.0±2.4	56.4±0.7	64.7±1.7	69.6±2.5	69.3±1.1	64.9±1.8	74.5±3.2	74.8±2.2	76.4±0.8	**76.5±1.6** (↑ 0.1)
	50%	55.9±1.9	56.5±1.4	66.3±2.2	69.3±2.4	68.0±1.5	63.9±2.9	73.1±1.1	73.9±0.9	75.5±1.9	**76.0±1.9** (↑ 0.5)
	70%	55.4±2.6	56.5±2.5	64.4±2.8	67.3±2.9	66.2±2.2	63.0±1.7	70.8±1.7	70.9±3.4	73.3±4.7	**74.5±3.0** (↑ 1.2)
	90%	55.4±2.7	56.3±2.4	63.7±2.7	65.9±3.2	62.3±3.9	61.9±3.1	69.2±3.1	68.8±3.4	71.6±4.7	**72.7±1.0** (↑ 1.1)
	Avg.	56.5	57.0	65.6	69.7	68.6	65.1	73.9	73.3	75.8	**76.1**
EYaleB& AR Light [Lighting]	0%	58.5	59.9	64.0	63.5	55.4	80.7	88.1	88.8	88.2	**90.6**
	10%	45.5±2.3	53.9±1.1	51.6±3.8	53.4±5.0	52.8±1.1	66.8±5.4	84.3±1.8	87.3±2.6	81.2±1.5	**89.8±1.0** (↑ 2.5)
	30%	37.9±1.7	50.5±2.1	48.3±4.8	53.1±5.5	46.5±2.3	61.4±3.8	81.3±3.4	83.4±2.4	79.5±5.0	**87.5±3.0** (↑ 4.1)
	50%	31.8±3.4	47.3±2.4	49.6±3.6	56.0±4.7	44.6±4.0	56.5±6.3	80.4±4.2	81.4±4.1	75.3±2.9	**88.1±1.7** (↑ 6.7)
	70%	26.1±2.1	40.1±1.7	47.2±2.2	54.6±2.5	38.0±2.7	53.3±3.0	75.5±2.6	77.8±2.2	71.8±3.9	**88.1±2.4** (↑ 10.3)
	90%	24.5±2.3	38.8±3.8	45.3±4.0	52.5±4.7	34.8±4.0	49.9±3.8	73.8±5.7	74.6±4.6	64.6±4.7	**88.0±4.0** (↑ 13.4)
	Avg.	37.4	48.4	51.0	55.5	45.4	61.4	80.6	82.2	76.8	**88.7**
CAS-PEAL [Disguise]	0%	51.3	51.4	62.3	69.5	59.2	75.8	78.7	78.2	80.3	**81.2**
	10%	45.0±1.0	45.3±0.9	54.7±1.4	61.4±2.7	54.4±0.6	73.1±0.9	76.7±1.2	76.5±1.3	77.0±3.3	**79.2±0.8** (↑ 2.2)
	30%	38.1±1.5	39.9±1.5	51.4±0.8	59.0±1.5	50.0±2.5	70.4±1.3	72.1±1.9	70.0±1.2	72.4±2.7	**76.3±0.7** (↑ 3.9)
	50%	32.3±1.4	35.2±2.0	47.1±2.1	56.3±2.7	44.5±2.5	66.2±0.5	67.9±2.1	65.9±1.2	67.6±2.3	**71.6±0.9** (↑ 3.7)
	70%	27.5±1.8	28.8±1.6	38.3±2.6	49.7±1.3	37.1±2.7	59.9±3.0	60.3±2.9	59.2±3.2	61.8±2.6	**67.2±1.8** (↑ 5.4)
	90%	23.8±1.8	29.1±2.0	39.7±2.5	49.8±1.8	35.1±1.8	59.1±3.7	59.6±2.4	57.3±1.1	60.0±3.5	**66.4±2.1** (↑ 6.4)
	Avg.	36.3	38.3	48.9	57.6	46.7	67.4	69.2	67.9	69.9	**73.7**
FERET [Pose]	0%	40.5	55.0	51.5	43.0	57.5	24.0	67.0	68.0	73.0	**90.5**
	10%	30.7±3.1	47.9±2.3	37.2±2.6	29.7±4.1	50.9±1.2	11.9±2.4	61.3±1.7	62.7±3.8	73.9±2.1	**89.6±0.8** (↑ 15.7)
	30%	24.7±3.6	37.4±3.1	30.4±2.0	25.8±2.6	41.8±1.4	10.8±1.4	53.5±1.7	59.1±3.2	70.3±3.9	**89.3±2.3** (↑ 19.0)
	50%	24.0±2.0	33.4±1.6	30.4±1.5	28.1±2.8	39.7±2.3	10.9±1.6	50.5±3.5	56.8±3.0	66.9±3.0	**88.0±0.5** (↑ 21.1)
	70%	24.6±1.5	32.2±1.2	29.5±0.9	28.7±1.0	36.8±2.7	12.3±1.4	47.0±1.7	54.5±2.7	63.6±1.2	**89.0±1.4** (↑ 25.4)
	90%	23.1±2.1	29.1±0.9	28.9±1.9	29.4±1.2	36.7±3.6	12.1±1.7	45.6±1.8	50.9±2.4	64.8±3.4	**90.0±0.8** (↑ 25.2)
	Avg.	27.9	39.2	34.7	30.8	43.9	13.7	54.2	58.7	68.8	**89.4**
AR [Multiple variations]	0%	42.2	44.9	49.6	50.8	51.9	74.1	76.0	76.6	77.8	**79.7**
	10%	35.7±2.2	39.3±1.1	42.3±4.4	43.4±4.8	46.4±1.9	67.7±3.7	70.6±0.8	71.9±1.8	77.0±2.6	**77.6±0.8** (↑ 0.6)
	30%	24.3±1.6	34.1±2.0	32.4±2.4	35.5±2.9	37.7±2.7	63.5±2.3	66.5±1.9	68.4±1.3	72.6±3.4	**75.1±1.4** (↑ 2.5)
	50%	23.0±1.1	30.0±2.7	31.2±2.1	34.7±2.0	32.4±1.2	56.1±1.4	61.2±1.4	61.7±2.7	69.4±3.0	**73.3±2.6** (↑ 3.9)
	70%	17.3±1.7	25.4±1.1	28.1±3.0	33.3±3.6	27.6±1.2	53.6±2.2	57.2±1.2	59.1±1.5	65.2±2.1	**69.3±2.0** (↑ 4.1)
	90%	14.3±0.8	23.3±1.3	25.6±1.4	31.0±1.9	24.0±1.6	48.9±2.5	53.5±0.8	55.1±4.1	58.0±4.6	**63.9±1.5** (↑ 5.9)
	Avg.	26.1	32.8	34.9	38.1	36.7	60.7	64.2	65.6	70.0	**73.2**

demonstrates that D^{id} plays the most important role in VD-GAN. This is because D^{id} is used to preserve the identity label, which contains the most important identity information. D^{gan} plays the second most important role because it is used to control the quality of the learned prototypes as well as to disentangle variations in the learned representations. G^{rec} is less important than the former two components since it only operates on the standard samples. D^{var} has the least importance among the four components. This can be explained by the fact that D^{gan} also targets at disentangling variations and thus may weaken the influence of D^{var} on the performance. We point out that we also have the same observations on the other four datasets and omit their results for conciseness. Furthermore, Fig. 8 shows the learned prototypes of an example input image by VD-GAN and the four VD-GAN variants on the AR dataset. We observed that, when removing D^{id}, the identity of the learned prototype is changed; when removing D^{gan}, the learned prototype has lower quality; when removing the reconstruction penalty G^{rec}, the learned prototype loses certain facial details; when removing D^{var}, the learned prototype looks close to that of VD-GAN, but still contains noises in the restored periocular area.



(a) VD-GAN w/o D^{id} vs. VD-GAN (b) VD-GAN w/o D^{var} vs. VD-GAN

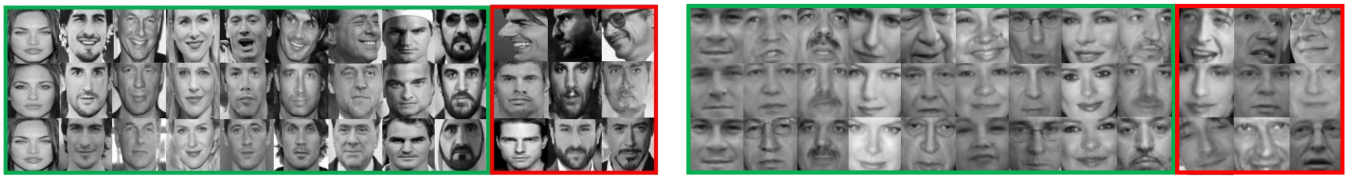
(c) VD-GAN w/o D^{gan} vs. VD-GAN (d) VD-GAN w/o G^{rec} vs. VD-GAN

Fig. 8. Examples of the learned prototypes on the AR dataset by VD-GAN and its four variants VD-GAN w/o D^{id}, VD-GAN w/o D^{var}, VD-GAN w/o D^{gan}, and VD-GAN w/o G^{rec}.

E. Evaluation on Mixed Variations Under Unconstrained Environment

In practice, it is possible that an enrolment sample is contaminated by complex mixed variations such as the combination of two or more different variations. In this subsection, we explore the feasibility of VD-GAN in such a challenging scenario. Specifically, we apply our VD-GAN to the unconstrained CFP and LFW datasets that contain mixed variations in the wild, and evaluate VD-GAN's performance for both SSPP-ce FR and prototype learning with contaminated enrolment samples. *To the best of our knowledge, this is the first work that learns prototypes for unconstrained faces with mixed variations.*

Following the setting in Subsection V-C, for each dataset, we construct different biometric enrolment databases with the contaminated ratio ranging from 0% to 50%. In the experiments, we use SRC as the baseline method and choose three


(a) CFP

(b) LFW

Fig. 9. Prototype learning on the unconstrained (a) CFP and (b) LFW datasets. Good examples are in the green box while the relatively bad ones are in the red box. The images from top to bottom lines are the original enrolment samples (the first one is standard sample and the rest are contaminated ones), the learned prototypes by VD-GAN, and the true prototypes for reference, respectively.

TABLE V

RANK-10 RECOGNITION RATES OF DIFFERENT METHODS FOR SSPP-CE FR ON CFP DATASET. THE CONTAMINATION RATIO IN THE ENROLMENT DATABASE RANGES FROM 0% TO 50%

Method	Recognition rate (%)			
	ratio=0%	ratio=10%	ratio=30%	ratio=50%
Baseline	44.3	42.8±0.2	41.6±0.6	39.2±0.9
SVDL	40.7	40.0±0.2	37.5±0.6	35.8±0.9
SLRC	46.8	42.6±0.1	40.2±1.0	38.1±1.4
S^{3}RC	46.0	42.8±0.2	40.4±0.9	37.8±2.1
VD-GAN	**66.1**	**63.2±0.7**	**62.1±1.2**	**58.5±1.0**

TABLE VI

RANK-10 RECOGNITION RATES OF DIFFERENT METHODS FOR SSPP-CE FR ON LFW DATASET. THE CONTAMINATION RATIO IN THE ENROLMENT DATABASE RANGES FROM 0% TO 50%

Method	Recognition rate (%)			
	ratio=0%	ratio=10%	ratio=30%	ratio=50%
Baseline	50.2	49.5±1.5	47.3±1.3	47.3±1.5
SVDL	57.6	55.5±1.9	54.1±1.7	52.9±1.6
SLRC	65.1	64.2±0.8	63.2±1.8	62.6±2.5
S^{3}RC	65.6	64.5±1.0	63.3±1.8	62.9±2.3
VD-GAN	**80.7**	**79.6±1.1**	**79.3±1.8**	**78.6±1.4**

other methods (i.e., SVDL, SLRC and S^{3}RC) that obtain the top-3 performance in Table IV for comparison. The parameters are set in the same way as in Subsection V-C. Table V and Table VI list the rank-10 recognition rates of different methods for SSPP-ce FR on the unconstrained CFP and LFW datasets, respectively. We have the following observations:

1) There exist large gaps between the performance in Table V-VI and that in Table IV, which demonstrates that it is rather challenging to perform SSPP-ce FR with mixed variations in the unconstrained setting.
2) VD-GAN consistently and significantly outperforms the other compared methods in all cases we have tried on both datasets. For example, on CFP, VD-GAN has a 19.3%, 20.4%, 20.5%, and 19.3% performance gain over the second best method as the contaminated ratio increases from 0% to 50%.
3) On CFP, state-of-the-art generic learning methods (i.e., SVDL and SLRC) achieve comparable or even worse performance compared to the baseline SRC. This indicates that existing generic learning methods are insufficient to handle mixed variations.

Furthermore, we visualize the learned prototypes of 12 randomly selected enrolment samples (including 1 standard sample) on CFP and LFW, respectively, in Fig. 9. We observe that our VD-GAN learns promising prototypes for a majority of the selected samples on both datasets. Particularly, for the standard samples and the samples with the mixed variations of slight poses and expressions, the learned prototypes by VD-GAN are almost the same as the true prototypes. Besides, VD-GAN also shows good capabilities to learn prototypes for the samples with the mixed variations of moderate poses and expressions, and small occlusions and lighting/expressions. However, in some extreme cases, e.g., mixed variations of large poses and expressions/occlusions, VD-GAN cannot generate satisfactory prototypes. One plausible reason is that key facial information is missing in these cases.

Note that some recent deep learning-based methods [7], [10], [65]–[68] have achieved promising performance for SSPP FR under unconstrained environments, by using the pre-trained models on large-scale web face datasets. Motivated by this, we further enhance the proposed VD-GAN by using the pre-trained LightCNN-29 feature extractor [69] on CASIA-WebFace [70] and MS-Celeb-1M [71] as the encoder G_{enc}. Moreover, we enforce the dimension of the extracted feature still be 320 by adding a FC layer (input: 256, output: 320) behind G_{enc}. The network structures of the decoder G_{dec} in G and the discriminator D are kept unchanged. In training, we freeze the parameters' values in the LightCNN-29 but update the parameters' values of the FC layer, G_{dec}, and D.

Subsequently, we evaluate VD-GAN using the LightCNN-29 feature extractor (i.e., VD-GAN$_{Lcnn}$) on LFW, and compare it with four recent deep learning-based approaches for SSPP FR including joint and collaborative representation with local adaptive convolution feature (JCR-ACF) [7], Regular-face [66], Arc-face [67], and the state-of-the-art class-level joint representation with regional adaptive convolution features (CJR-RACF) [10]. We follow the evaluation protocol suggested in JCR-ACF, and report the rank-1 recognition rates of all the methods for SSPP FR in Table VII. It can be observed that our VD-GAN$_{Lcnn}$ achieves a promising recognition rate of 98.4%, which is higher than that of the other four compared deep learning-based methods.

In general, the experimental results in Fig. 9 and Table V-VI have shown the effectiveness of our VD-GAN to learn prototypes for in-the-wild faces containing mixed variations, as well as the superiority to learn representations for solving SSPP-ce FR over the existing generic learning methods. Moreover, the inspiring recognition result of VD-GAN$_{Lcnn}$ on LFW in Table VII verifies the feasibility of combining our VD-GAN

TABLE VII

RANK-1 RECOGNITION RATES (%) OF VD-GAN USING THE LIGHT-CNN FEATURE ENCODER, I.E., VD-GAN$_{Lcnn}$, AND THE OTHER DEEP LEARNING-BASED METHODS ON LFW DATASET

Methods	Recognition rate (%)
JCR-ACF	86.0
Regular-face	83.7
Arc-face	92.3
CJR-RACF	95.5
VD-GAN$_{Lcnn}$	**98.4**

with pre-trained deep feature extractors and provides a new promising direction for solving practical SSPP FR.

VI. CONCLUSION

We have proposed the VD-GAN model, which is the first attempt to jointly learn prototypes and representations from the contaminated SSPP. VD-GAN is able to deal with universal variations, including specified single variation, unspecified multiple variations, and even mixed variations, in the biometric enrolment database. The proposed VD-GAN consists of an encoder-decoder structural generator and a multi-task discriminator, which play an adversarial game such that 1) the learned prototype of each enrolment sample (i.e., identity) can recover his/her standard face, and 2) the learned discriminative and variation-free representations for enrolment samples and query samples can be used to perform the challenging SSPP-ce FR. Extensive experiments on various real-world face datasets containing single/multiple and mixed variations have demonstrated the effectiveness of VD-GAN for joint prototype learning and representation learning. Furthermore, to enhance the representation learning ability of VD-GAN under unconstrained environments, it is feasible to employ a powerful deep feature extractor pre-trained on large-scale web face datasets as the encoder in the generator.

REFERENCES

[1] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, Sep. 2006.

[2] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014.

[3] S. Gao, K. Jia, L. Zhuang, and Y. Ma, "Neither global nor local: Regularized patch-based representation for single sample per person face recognition," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 365–383, Feb. 2015.

[4] G. Zhang, H. Sun, Z. Ji, Y.-H. Yuan, and Q. Sun, "Cost-sensitive dictionary learning for face recognition," *Pattern Recognit.*, vol. 60, pp. 613–629, Dec. 2016.

[5] G. Zhang, H. Sun, Z. Ji, and Q. Sun, "Label propagation based on collaborative representation for face recognition," *Neurocomputing*, vol. 171, pp. 1193–1204, Jan. 2016.

[6] Z.-M. Li, Z.-H. Huang, and K. Shang, "A customized sparse representation model with mixed norm for undersampled face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2203–2214, Oct. 2016.

[7] M. Yang, X. Wang, G. Zeng, and L. Shen, "Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person," *Pattern Recognit.*, vol. 66, pp. 117–128, Jun. 2017.

[8] F. Mokhayeri, E. Granger, and G.-A. Bilodeau, "Domain-specific face synthesis for video face recognition from a single sample per person," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 3, pp. 757–772, Mar. 2019.

[9] M. Pang, Y.-M. Cheung, B. Wang, and J. Lou, "Synergistic generic learning for face recognition from a contaminated single sample per person," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 195–209, 2020.

[10] M. Yang, W. Wen, X. Wang, L. Shen, and G. Gao, "Adaptive convolution local and global learning for class-level joint representation of facial recognition with a single sample per data subject," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2469–2484, 2020.

[11] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2021.

[12] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[14] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 1713–1726.

[15] Y. Zhou and S. Sun, "Manifold partition discriminant analysis," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 830–840, Apr. 2017.

[16] M. Pang, B. Wang, Y.-M. Cheung, and C. Lin, "Discriminant manifold learning via sparse coding for robust feature extraction," *IEEE Access*, vol. 5, pp. 13978–13991, 2017.

[17] M. Pang, Y.-M. Cheung, R. Liu, J. Lou, and C. Lin, "Toward efficient image representation: Sparse concept discriminant matrix factorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3184–3198, Nov. 2019.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[19] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 471–478.

[20] T. Pei, L. Zhang, B. Wang, F. Li, and Z. Zhang, "Decision pyramid classifier for face recognition under complex variations using single sample per person," *Pattern Recognit.*, vol. 64, pp. 305–313, Apr. 2017.

[21] P. Zhu, L. Zhang, Q. Hu, and S. C. Shiu, "Multi-scale patch based collaborative representation for face recognition with margin distribution optimization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 822–835.

[22] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, Jan. 2013.

[23] F. Liu, J. Tang, Y. Song, L. Zhang, and Z. Tang, "Local structure-based sparse representation for face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 1, pp. 1–20, Oct. 2015.

[24] P. Zhang, X. You, W. Ou, C. L. P. Chen, and Y.-M. Cheung, "Sparse discriminative multi-manifold embedding for one-sample face identification," *Pattern Recognit.*, vol. 52, pp. 249–259, Apr. 2016.

[25] 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.

[26] C.-P. Wei and Y.-C.-F. Wang, "Undersampled face recognition via robust auxiliary dictionary learning," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1722–1734, Jun. 2015.

[27] W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 399–406.

[28] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2513–2521, Oct. 2018.

[29] M. Yang, L. Van, and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 689–696.

[30] H.-K. Ji, Q.-S. Sun, Z.-X. Ji, Y.-H. Yuan, and G.-Q. Zhang, "Collaborative probabilistic labels for face recognition from single sample per person," *Pattern Recognit.*, vol. 62, pp. 125–134, Feb. 2017.

[31] 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.

[32] 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.*, early access, Apr. 20, 2020, doi: 10.1109/TNNLS.2020.2985099.

[33] W. Ma, X. Xie, C. Yin, and J. Lai, "Face image illumination processing based on generative adversarial nets," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2558–2563.

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

[35] 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.

[36] 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.

[37] Y. Zhao *et al.*, "Identity preserving face completion for large ocular region occlusion," 2018, *arXiv:1807.08772*. [Online]. Available: http://arxiv.org/abs/1807.08772

[38] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3990–3999.

[39] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[40] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang, "Single sample face recognition via learning deep supervised autoencoders," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 10, pp. 2108–2118, Oct. 2015.

[41] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[42] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2539–2547.

[43] 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.

[44] 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.

[45] X. Wang and X. Tang, "Bayesian face recognition based on Gaussian mixture models," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 142–145.

[46] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Robust semi-supervised subspace clustering via non-negative low-rank representation," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1828–1838, Aug. 2016.

[47] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[48] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1991, pp. 586–587.

[49] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[50] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2143–2156, Dec. 2007.

[51] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by nuclear/l2,1-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, Apr. 2017.

[52] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1558–1566.

[53] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[54] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[55] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[56] W. Gao *et al.*, "The CAS-PEAL large-scale chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[57] 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, 2000.

[58] A. Martinez and R. Benavente, "The AR face database," Comput. Vis. Center, Barcelona, Spain, Tech. Rep. 24, Jun. 1998.

[59] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[60] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images, Detection, Alignment, Recognit.*, 2008, pp. 1–15.

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

[62] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[63] S. Robertson, "A new interpretation of average precision," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 689–690.

[64] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 144–159, 2020.

[65] O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, vol. 1, no. 3, p. 6.

[66] K. Zhao, J. Xu, and M.-M. Cheng, "RegularFace: Deep face recognition via exclusive regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1136–1144.

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

[68] V. Cuculo, A. D'Amelio, G. Grossi, R. Lanzarotti, and J. Lin, "Robust single-sample face recognition by sparsity-driven sub-dictionary learning using deep features," *Sensors*, vol. 19, no. 1, p. 146, Jan. 2019.

[69] 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.

[70] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

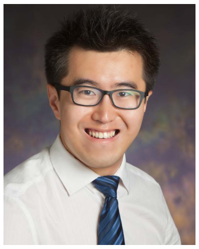
[71] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 87–102.

Meng Pang received the B.Sc. and M.Sc. degrees in software engineering from the 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, in 2019. He is currently a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include image processing and adversarial machine learning.

Binghui Wang (Member, IEEE) received the B.Sc. degree in network engineering and the M.Sc. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical and computer engineering from Iowa State University, Ames, Iowa, in 2019. He is currently a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Duke University. His research interests include data-driven security and privacy, trustworthy machine learning, and machine learning.

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. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China. His research interests include machine learning, pattern recognition, visual computing, and optimization. He is a fellow of IET, BCS, and RSA, and an IETI Distinguished Fellow. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, and *Pattern Recognition*, to name a few.

Yiran Chen (Fellow, IEEE) received the B.S. and M.S. degrees from Tsinghua University in 1998 and 2001, respectively, and the Ph.D. degree from Purdue University in 2005. After five years in industry, he joined the University of Pittsburgh in 2010 as an Assistant Professor and then promoted to an Associate Professor with tenure in 2014, held a Bicentennial Alumni Faculty Fellow. He is currently a Professor with the Department of Electrical and Computer Engineering, Duke University, and serving as the Director of the NSF Industry-University Cooperative Research Center (IUCRC) for Alternative Sustainable and Intelligent Computing (ASIC) and the Co-Director of the Duke Center for Computational Evolutionary Intelligence (CEI), focusing on the research of new memory and storage systems, machine learning and neuromorphic computing, and mobile computing systems. He has published one book and more than 400 technical publications and has been granted 96 U.S. patents. He is a fellow of ACM. He received seven best paper awards, one Best Poster Award, and fourteen best paper nominations from international conferences and workshops. He was a recipient of the NSF CAREER Award, the ACM SIGDA Outstanding New Faculty Award, the Humboldt Research Fellowship for Experienced Researchers, and the IEEE SYSC/CEDA TCCPS Mid-Career Award. He serves or served as an Associate Editor for more than ten international academic transactions/journals and served on the technical and organization committees for more than 60 international conferences. He is also serving as the Editor-in-Chief for the *IEEE Circuits and Systems Magazine*. He is a Distinguished Lecturer of IEEE CEDA and listed in the HPCA Hall of Fame.



Bihan Wen (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, USA, in 2015 and 2018, respectively. He was a Researcher with Dolby Laboratories, CA, USA. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computer vision, image and video processing, computational imaging, and big data applications. He 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. His coauthored paper received the Top 10% Best Paper Award at the IEEE International Conference on Image Processing in 2014 and another received the Best Paper Runner-Up Award at the IEEE International Conference on Multimedia and Expo in 2020.