DisP+V: A Unified Framework for Disentangling Prototype and Variation From Single Sample per Person

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

***Abstract*— Single sample per person face recognition (SSPP FR) is one of the most challenging problems in FR due to the extreme lack of enrolment data. To date, the most popular SSPP FR methods are the generic learning methods, which recognize query face images based on the so-called prototype plus variation (i.e., P+V) model. However, the classic P+V model suffers from two major limitations: 1) it *linearly* combines the prototype and variation images in the observational pixel-spatial space and cannot generalize to multiple *nonlinear* variations, e.g., poses, which are common in face images and 2) it would be severely impaired once the enrolment face images are contaminated by nuisance variations. To address the two limitations, it is desirable to disentangle the prototype and variation in a latent feature space and to manipulate the images in a semantic manner. To this end, we propose a novel *disentangled* prototype plus variation model, dubbed DisP+V, which consists of an encoder–decoder generator and two discriminators. The generator and discriminators play two adversarial games such that the generator nonlinearly encodes the images into a latent semantic space, where the more discriminative prototype feature and the less discriminative variation feature are disentangled. Meanwhile, the prototype and variation features can guide the generator to generate an identity-preserved prototype and the corresponding variation, respectively. Experiments on various real-world face datasets demonstrate the superiority of our DisP+V model over the classic P+V model for SSPP FR. Furthermore, DisP+V demonstrates its unique characteristics in both prototype recovery and face editing/interpolation.**

***Index Terms*— Adversarial learning, disentangled representation, face editing, prototype recovery, single sample per person.**

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 several important real-world applications, such as criminal identification, surveillance security, access control, and person re-identification [1]–[15]. SSPP FR is still one of the most challenging problems in FR due to the extreme lack of enrolment data and the unavailability of intraclass information [16]. In such a case, a flurry of popular Fisher-based methods [17]–[21] are typically inapplicable. Moreover, many existing sparse representation and dictionary learning methods [22]–[25] will also suffer serious performance drop because they require sufficient samples to represent query samples.

To date, the most studied SSPP FR methods are the generic learning methods [26]–[31], which are based on a so-called prototype plus variation (i.e., P+V) model for recognition. In the P+V model, a query sample is assumed to be represented by the superposition of the prototype[2] and the corresponding facial variations [32]. The prototype is approximated by the original enrolment sample, while the variation dictionary is generated from an auxiliary generic set that encodes the difference between the query and enrolment samples. The major differences between these generic learning methods lie in the strategies of learning the variation dictionary.

However, the classic P+V model has two major limitations. First, it is a *linear* model that combines the prototype and variation images in the pixel-spatial space, which is unable to handle many nonlinear variations, such as poses. Despite that, the classic P+V model ignores the importance of different components (e.g., eyes, nose, and cheeks) in the face image and assigns them the same weights when performing combination. In Fig. 1(a), we show a failed reconstruction example of a state-of-the-art generic learning method, i.e., superposed

Manuscript received 6 December 2020; revised 14 May 2021; accepted 31 July 2021. Date of publication 17 August 2021; date of current version 6 February 2023. The work of Meng Pang and Bihan Wen was supported in part by the Ministry of Education, Singapore, under a startup grant, in part by the National Research Foundation Singapore, and in part by the Singapore Cybersecurity Consortium (SGCSC) Grant Office under Grant SGCSC_Grant_2019-S01. The work of Yiu-ming Cheung was supported in part by NSFC under Grant 61672444; in part by Hong Kong Baptist University under Grant RC-FNRA-IG/18-19/SCI/03, Grant RC-IRCMs/18-19/SCI/01, and Grant AIS 21-22/03; and in part by the Innovation and Technology Fund of the Innovation and Technology Commission, Government of the Hong Kong SAR, under Project ITS/339/18. *(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 is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: bwang70@iit.edu).

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

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

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3103194.

Digital Object Identifier 10.1109/TNNLS.2021.3103194

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

[2]A prototype indicates a frontal face image with a neutral expression, under normal lighting, and without occlusion/disguise.

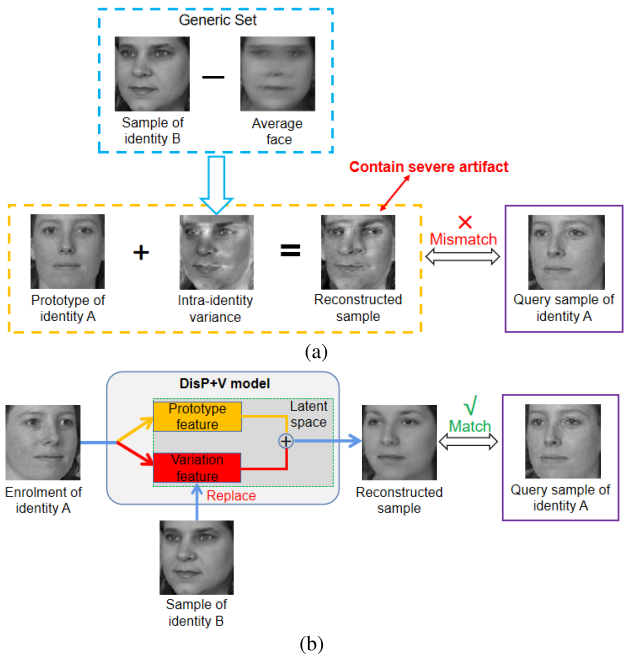


Fig. 1. (a) Failed reconstruction example of the classic P+V model-based SLRC [26] when dealing with poses. In the classic P+V model, the prototype of identity A and the generated pose variation from the generic set is superposed in the spatial space to reconstruct the query sample. (b) Illustration of our DisP+V model, where the prototype and variation features of the enrolment sample are disentangled in the latent space. We replace the variation feature with the one disentangled from the sample of identity B with the target pose and perform the superposition of the prototype and variation in a semantic manner.

linear representation classifier (SLRC) [26], when dealing with the pose variation. It is observed that the reconstructed image contains obvious artifacts and individual characteristics of the other identity if the generated pose variation is simply integrated into the prototype. Second, the P+V model estimates the prototype by directly using the *standard* enrolment sample. When facing the emerging problem in SSPP FR, namely, SSPP FR with a contaminated enrolment database[3] (i.e., SSPP-ce FR) [33], the P+V model will be severely impaired because the *contaminated* enrolment sample yields bad prototype to represent the identity. For clarity, the previous problem of SSPP FR with a standard enrolment database is called SSPP-se FR hereinafter.

To address these limitations, it is desired to seek a latent semantic space where the more discriminative prototype feature and the less discriminative variation feature can be successfully disentangled. Subsequently, the prototype feature for the enrolment sample and the variation feature for the generic sample are obtained, and the feature of the query sample is approximated by the superposition of the prototype and variation features in this latent space. To this end, we propose a novel *disentangled* prototype plus variation (DisP+V) model for SSPP FR, as shown in Fig. 1(b). Compared with the classic P+V model that linearly combines the prototype and variation images in the pixel-spatial space, our proposed DisP+V model is featured in two aspects.

1) It is a top-down P+V model that performs the combination of prototype and variation in a latent semantic space, which could *implicitly* lead to an adaptive weighting of different image components in the pixel-spatial space.
2) It results in better discrimination between the prototype and variation by mining the underlying properties.

The advantages make DisP+V capable of handling both linear and nonlinear variations. Moreover, DisP+V is robust against the enrolment contaminations in SSPP-ce FR, because it first extracts the discriminative prototype feature from the contaminated enrolment sample and then performs the superposition of prototype feature and variation feature in the latent space.

To be specific, DisP+V consists of three main components, i.e., an encoder–decoder structural generator (G) and two discriminators $D = [D^{\text{id}}, D^{\text{gan}}]$ and $\tilde{D}$, where D^{id} and $\tilde{D}$ are used for predicting face identity and D^{gan} for distinguishing real versus fake prototype. Fig. 2 shows the architecture of the proposed DisP+V model. Given an input face, the three components G, D, and $\tilde{D}$ play two adversarial games: 1) G strives for generating an identity-preserved prototype to fool D, while D guides G to encode a discriminative prototype feature relevant to identity and 2) G and $\tilde{D}$ compete with each other such that G encodes a less discriminative variation feature, and meanwhile, generating the corresponding variation image that fools $\tilde{D}$, i.e., G enables $\tilde{D}$ to output a constant vector with a uniform identity distribution. Furthermore, DisP+V introduces a reconstruction penalty in G to force the decoded image from the superposition of the prototype and variation features to well reconstruct the input face, which guarantees the complementarity between the two disentangled features.

We conduct experiments on six real-world face datasets containing a single variation of expression, pose, disguise, and lighting, multiple variations, and mixed variations in the wild, respectively. Our experimental results demonstrate the superiority of the proposed DisP+V model over the classic P+V model for both SSPP-se FR and SSPP-ce FR. For instance, on Face Recognition Technology (FERET) dataset [34], DisP+V achieves a 30.1%–39.9% higher accuracies than the state-of-the-art P+V-based generic learning method for SSPP-ce FR. Moreover, note that recent deep learning-based methods [5], [11], [35]–[37] have achieved promising performance for practical SSPP FR benefiting from the pretrained models on large-scale Web face datasets. Motivated by this, we, thus, enhance DisP+V by employing a pretrained deep feature extractor as the encoder and verify the feasibility and effectiveness of this combination in the experiments. Furthermore, DisP+V has demonstrated its unique characteristics for handling challenging tasks of prototype recovery and face editing/interpolation.

To the best of our knowledge, the proposed DisP+V is the first attempt that jointly: 1) disentangles the prototype and variation features in the latent space and 2) generates the corresponding prototype and variation image, in a unified deep framework. Moreover, DisP+V only constrains the low discriminative property of the disentangled variation but has no prior assumption about its type, which makes DisP+V applicable to universal variations. The contributions of this article are summarized as follows.

1) We propose DisP+V, a top-down P+V model for solving SSPP FR. Compared with the classic P+V model that can only deal with linear variations and

[3]A contaminated enrolment database means that some face samples in the enrolment database are contaminated by different facial variations.

standard enrolment, DisP+V is effective in handling both linear and nonlinear variations and the enrolment contaminations.

2) We design an encoder–decoder structural generator in DisP+V that can simultaneously: 1) learn the prototype and variation features and 2) generate the corresponding prototype and variation images, from a contaminated enrolment sample.
3) We design two adversarial discriminators to assist the generator in: 1) removing the variations and meanwhile preserving the identity information of the input contaminated enrolment sample in the generated prototype and its feature and 2) eliminating the identity information in the generated variation and its feature.
4) We conduct extensive experiments on various real-world face datasets with single/multiple and mixed variations to demonstrate the powerful capability of DisP+V for prototype recovery and face editing (or interpolation) and the superiority for SSPP FR over the classic P+V model-based counterparts.

The rest of this article is organized as follows. Section II makes an overview of the related works, and Section III gives a review of the classic P+V model and the generative adversarial network (GAN). Section IV details the proposed DisP+V. In Section V, we perform extensive experiments on six real-world face datasets to evaluate the performance of DisP+V. Finally, Section VI gives the conclusion and future works.

II. Related Work

A. SSPP FR

In the past decade, many attempts have been made for solving the SSPP-se FR problem, where all enrolment samples are standard, which can be roughly classified into two categories [38], i.e., patch-based methods and generic learning methods.

The patch-based methods [39]–[42] partition each enrolment sample into multiple local patches and then leverage them for discriminative learning or feature extraction. However, the local patches from a single sample contain limited and highly correlated information which are hardly treated as independent samples. By introducing new and useful information from the auxiliary generic set, the generic learning methods [26], [28]–[30] usually perform better than the patch-based methods and receive more attention. These methods generate the variation dictionaries from the generic set and utilize the classic P+V model [32] for recognition. For example, Deng *et al.* [26] generate the variation dictionary by subtracting the average face from the samples of each identity in the generic set, while Yang *et al.* [28] propose to project the generic set into the space of enrolment set and learn an adaptive sparse variation dictionary. However, the P+V model used in these methods is a simple linear superposition model and can hardly handle nonlinear variations. Despite that, the prototype in the P+V model is directly estimated by the original enrolment samples, which makes the existing generic learning methods not amenable to tackle enrolment contaminations.

More recently, a few prototype learning methods [8], [43]–[47] have been proposed to address the new SSPP-ce FR problem, where some enrolment samples can be contaminated. Gao *et al*. [43] and Pang *et al*. [8] proposed a semisupervised sparse representation-based classification (S^3RC) and an iterative dynamic generic learning (IDGL), respectively. The two methods estimate the prototypes by the clustering centroid of the union of enrolment and query sets via the Gaussian mixture model (GMM) or semisupervised low-rank representation. Despite promising prototypes obtained by S^3RC and IDGL, they need to obtain the unknown query set in advance, which is difficult to satisfy in practice. Furthermore, a series of GAN variants [44]–[47] emerge to recover prototypes by virtue of adversarial learning. For example, Ma *et al*. [44] proposed a style translation GAN to learn the mappings between arbitrary lighting domains and standard lighting domain for normalization; Huang *et al*. [47] 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 or pose, they need to know the input type of the variation in advance and cannot handle unspecified multiple variations.

B. Face Disentangled Representation

Face disentangled representation is a kind of distributed feature representation where different latent codes reflect different high-level generative factors of the face image, such as ID-related feature map, facial attributes or variations, and artistic style. Kingma and Welling [48] developed a variational auto-encoder (VAE) to disentangle the factors of variation and learn the latent code by encouraging the latent distribution to be close to the standard normal distribution. Larsen *et al.* [49] extended VAE by employing a learned similarity measure in GAN discriminator as the reconstruction objective instead of the original elementwise residual. Liu *et al*. [50] presented an identity distilling and dispelling AE to learn the identity-distilled feature for identity verification and the identity-dispelled features to fool the verification system. Although these AE-based methods can be applied for solving the SSPP-ce FR problem, they are unable to perform prototype recovery tasks at the same time. Lately, Kulkarni *et al*. [51] proposed a deep convolution inverse graphics network to generate representations disentangled w.r.t. pose or lighting. Tran *et al*. [52], [53] proposed a disentangled representation learning GAN, which learns a pose-invariant representation and meanwhile rotating input face to a specified pose. These two methods can perform pose frontalization while learning disentanglement representations. However, both of them are limited to representing a specified single variation and cannot generalize to multiple variations. In contrast to the above-mentioned approaches, our DisP+V jointly: 1) disentangles the prototype and variation features in the latent space and 2) generates the corresponding prototype and variation images and is able to handle universal variations.

III. Background

A. Prototype Plus Variation Model

The classic prototype plus variation (i.e., P+V) model [32] is developed to handle the SSPP-se FR problem, which is based on the assumption that a query sample of one identity

is represented as a superposition of two different subsignals, i.e., the prototype of the identity plus the intra-identity variations. In 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. Formally, for a query 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 small noise, respectively, $\boldsymbol{\alpha}$ is the sparse coefficient vector, whose a few nonzero entries correspond to choosing a few numbers of enrolment samples (i.e., identities) from $\mathbf{P}$, and $\boldsymbol{\beta}$ is another sparse coefficient vector whose nonzero entries correspond to selecting a small subset of dictionary $\mathbf{V}$. The coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are calculated via solving the following optimization problem:

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

where λ is a regularization parameter, $||.||_2$ and $||.||_1$ denote the l_2-norm and l_1-norm, respectively. Finally, similar to sparse representation-based classification (SRC) [22], $\mathbf{y}$ will be classified into the enrolment sample with the smallest reconstruction residual. Note that, the classic P+V model is a linear superposition model that manipulates images in the pixel-spatial space, which is difficult to process the nonlinear variations. Moreover, when confronting the more challenging SSPP-ce FR problem where enrolment samples are contaminated, this model will be severely impaired because the contaminated samples yield bad prototypes to represent the identities.

B. Generative Adversarial Network

Goodfellow *et al.* [54] proposed the GAN to train a generative model. It is composed of two 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 fake generated image $\hat{\mathbf{x}}$, while the generator G is trained to generate realistic-looking images, i.e., $G(\mathbf{z})$, based on a random noise vector $\mathbf{z}$ to fool D. Formally, the objective function of GAN is as follows:

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim\mathcal{P}_{\text{data}}}\left[\log D(\mathbf{x})\right] + E_{\mathbf{z}\sim\mathcal{P}_z}\left[\log(1 - D(G(\mathbf{z})))\right] \tag{3}$$

where p_{data} and p_z denote the distributions of the training data and the noise $\mathbf{z}$, respectively. Alternatively, it has been shown 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 [54]. Hence, the objective in (3) can be reformulated as follows:

$$\max_D V_D(G,D) = E_{\mathbf{x}\sim\mathcal{P}_{\text{data}}}\left[\log D(\mathbf{x})\right] + E_{\mathbf{z}\sim\mathcal{P}_z}\left[\log(1 - D(G(\mathbf{z})))\right], \tag{4}$$

$$\max_G V_G(G,D) = E_{\mathbf{z}\sim\mathcal{P}_z}\left[\log(D(G(\mathbf{z})))\right]. \tag{5}$$

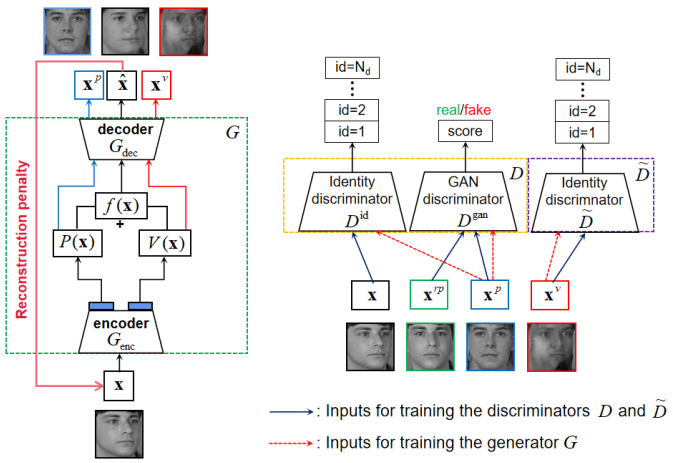


Fig. 2. Architecture of the proposed DisP+V. $\mathbf{x}$, $\mathbf{x}^p$, $\mathbf{x}^v$, $\hat{\mathbf{x}}$, and $\mathbf{x}^{\text{rp}}$ are the input face, the generated prototype, the generated variation, the reconstructed face, and the real prototype, respectively. $P(\mathbf{x})$ and $V(\mathbf{x})$ are the disentangled prototype and variation features in the encoded latent space, respectively, and $f(\mathbf{x}) = P(\mathbf{x}) + V(\mathbf{x})$. When training D, D^{id} predicts the identity label of $\mathbf{x}$, and D^{gan} assigns a high score to the real prototype $\mathbf{x}^{\text{rp}}$ but a low score to the generated prototype $\mathbf{x}^p$. When training $\tilde{D}$, $\tilde{D}$ predicts the identity of $\mathbf{x}^v$. When training G, the generated prototype $\mathbf{x}^p$ aims to fool D^{id} and D^{gan} to classify it into the identity label of $\mathbf{x}$ and to assign it a high score of being real prototype, respectively; the reconstructed face $\hat{\mathbf{x}}$ aims to well reconstruct the input face $\mathbf{x}$; and the generated variation $\mathbf{x}^v$ aims to fool $\tilde{D}$ to output a constant vector with a uniform distribution.

IV. Proposed Method

In this section, we first define the problem we are addressing. Then, we detail the proposed *disentangled* prototype plus variation (DisP+V) with the network architecture and training scheme. Finally, we introduce the potential applications.

A. Problem Definition

We propose a top-down P+V model which performs the superposition of the prototype and variation in a latent space, thus manipulating the images in a semantic manner without the complex combination designs in the pixel-spatial space. This so-called DisP+V model aims to learn disentangled prototype and variation features and to generate an identity-preserved prototype and the corresponding variation image.

To be specific, given an input face image $\mathbf{x}$, the proposed DisP+V aims to achieve the following objectives.

1) *Disentangled Feature Learning:* Learning a discriminative prototype feature $P(\mathbf{x})$ for $\mathbf{x}$, such that $P(\mathbf{x})$: 1) represents the identity of $\mathbf{x}$ and 2) is invariant to any facial variations in $\mathbf{x}$; and learning a less discriminative variation feature $V(\mathbf{x})$ for $\mathbf{x}$ such that $V(\mathbf{x})$ is irrelevant to the input identity information.
2) *Prototype and Variation Generation:* Recovering a high-quality (i.e., realistic looking) prototype $\mathbf{x}^p$ for the input face image $\mathbf{x}$, such that $\mathbf{x}^p$: 1) is variation-free and 2) preserves the identity of $\mathbf{x}$; and extracting the variation image $\mathbf{x}^v$ such that it: 1) captures the facial variation in $\mathbf{x}$ and 2) contains little identity information of $\mathbf{x}$.

B. DisP+V

In this section, we introduce the proposed *disentangled* prototype plus variation (DisP+V) model, whose architecture is shown in Fig. 2. The proposed DisP+V consists of three

TABLE I
MEANING OF THE SYMBOLS IN DISP+V

Symbol	Definition
$\mathbf{x}$	Image in training set, $\mathbf{x} \sim \mathcal{P}_{data}$
l^{id}/l^{var}	The identity/variation label for $\mathbf{x}$
$\mathbf{x}^{rp}$	Real prototype in training set, $\mathbf{x}^{rp} \sim \mathcal{P}_{real}$
G	The encoder-decoder structural generator
G_{enc}	The encoder in G
$P(\mathbf{x})$	The learned prototype feature of $\mathbf{x}$
$V(\mathbf{x})$	The learned variation feature of $\mathbf{x}$
$f(\mathbf{x})$	The superposition of $P(\mathbf{x})$ and $V(\mathbf{x})$
G_{dec}	The decoder in G
$\mathbf{x}^p$	The generated prototype of $\mathbf{x}$
$\mathbf{x}^v$	The generated variation image of $\mathbf{x}$
$\widehat{\mathbf{x}}$	The reconstructed image of $\mathbf{x}$
D	The multi-task discriminator, i.e., $D = [D^{gan}, D^{id}]$
D^{gan}	The sub-discriminator to classify real and fake prototypes
D^{id}	The sub-discriminator to predict face identity
$\tilde{D}$	The identity discriminator only relates to $\mathbf{x}^v$
$\mathbf{y}$	Query image in testing set
$\mathbf{S}$	The SSPP enrolment set in testing set
$P(\mathbf{y})$	The learned prototype feature of $\mathbf{y}$
$P(\mathbf{S})$	The learned prototype features of $\mathbf{S}$

main parts: an encoder–decoder structural network serving as the generator G, and two discriminators D and $\tilde{D}$ for adversarial learning. In the following, we will detail the generator G and the two discriminators D and $\tilde{D}$, followed by the training and evaluation schemes. Table I summarizes the symbols and the corresponding definitions used in DisP+V.

1) Generator and Discriminators: 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} has two separate branches, which aim to encode a more discriminative prototype feature $P(\mathbf{x})$ and a less discriminative variation feature $V(\mathbf{x})$ in a latent space. Subsequently, G_{dec} takes $P(\mathbf{x})$, $V(\mathbf{x})$ and their superposition, i.e., $f(\mathbf{x}) = P(\mathbf{x}) + V(\mathbf{x})$, as the inputs, and generates an appropriate prototype, i.e., $\mathbf{x}^p = G_{\text{dec}}(P(\mathbf{x}))$, a variation image, i.e., $\mathbf{x}^v = G_{\text{dec}}(V(\mathbf{x}))$, and a reconstructed image of $\mathbf{x}$, i.e., $\widehat{\mathbf{x}} = G_{\text{dec}}(f(\mathbf{x}))$, respectively.

The proposed D is a multitask discriminator consisting of two subdiscriminators, namely, D^{id} and D^{gan}. To be specific, the following holds.

1) D^{id} outputs a N_d-dimensional vector for face identity classification, with N_d the total number of identities.
2) D^{gan} is a standard GAN discriminator to distinguish the real prototype versus 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.

The proposed $\tilde{D}$ is also an identity discriminator that outputs a N_d-dimensional vector and is used to predict the face identity label. Unlike D^{id} of D, $\tilde{D}$ only relates to the variation image $\mathbf{x}^v$.

2) DisP+V Training: Suppose we are given a training set of N_d identities with each face image $\mathbf{x}$ annotated by the label $l = \{l^{\text{id}}, l^{\text{var}}\}$, where l^{id} and l^{var} ($l^{\text{var}} = 1$ or 0) denote the face identity and whether the face contains variation or not, respectively. Subsequently, we collect standard images (i.e., images not corrupted by variations) in the training set according to the l^{var} to form the real prototype corpus. We denote each standard/real prototype as $\mathbf{x}^{\text{rp}}$, and its distribution as $\mathcal{P}_{\text{real}}$, i.e., $\mathbf{x}^{\text{rp}} \sim \mathcal{P}_{\text{real}}$. As a comparison, we denote that all face images $\mathbf{x}$ in the training set are sampled from the distribution $\mathcal{P}_{\text{data}}$, i.e., $\mathbf{x} \sim \mathcal{P}_{\text{data}}$.

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

1) Enable D^{id} to classify the generated prototype $\mathbf{x}^p$ as the same identity label as the input image $\mathbf{x}$, i.e., l^{id}.
2) Fool D^{gan} to classify the generated *fake* prototype $\mathbf{x}^p$ as a real prototype, i.e., G enables D^{gan} to assign a high score to $\mathbf{x}^p$ of being real prototype.
3) Fool $\tilde{D}$ and make it fail to classify the generated variation $\mathbf{x}^v$, i.e., G enables $\tilde{D}$ to output a constant vector with each element value equaling to $(1/N_d)$.
4) Enable $\widehat{\mathbf{x}}$ to well reconstruct the original input image $\mathbf{x}$.

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

$$\max_G V_G = V_G^{\text{gan}} + \mu_1 V_G^{\text{id1}} + \mu_2 V_G^{\text{id2}} - \mu_3 V_G^{\text{rec}} \tag{6}$$

where μ_1, μ_2, and μ_3 are three positive tradeoff parameters for the hybrid objective V_G. The four subobjectives V_G^{id1}, V_G^{gan}, V_G^{id2}, and V_G^{rec} are defined as follows:

$$V_G^{\text{id1}}(G, D^{\text{id}}, \mathbf{x}) = E_{\mathbf{x}}\left[\log D_{l^{\text{id}}}^{\text{id}}(G_{\text{dec}}(P(\mathbf{x})))\right] \tag{7}$$

$$V_G^{\text{gan}}(G, D^{\text{gan}}, \mathbf{x}) = E_{\mathbf{x}}\left[\log D^{\text{gan}}(G_{\text{dec}}(P(\mathbf{x})))\right] \tag{8}$$

$$V_G^{\text{id2}}(G, \tilde{D}, \mathbf{x}) = H_{\mathbf{x}}\left[\tilde{D}(G_{\text{dec}}(V(\mathbf{x})))\right] \tag{9}$$

$$V_G^{\text{rec}}(G, \mathbf{x}) = E_{\mathbf{x}}\left[\frac{1}{2}||\mathbf{x} - G_{\text{dec}}(f(\mathbf{x}))||_F^2\right] \tag{10}$$

where D_i^{id} denotes the ith element in D^{id}, $H(.)$ and $||.||_F$ denote the empirical entropy and the Frobenius norm, respectively. It is worth noting that maximizing entropy of the predicted identity distribution for $G_{\text{dec}}(V(\mathbf{x}))$, i.e., $H_{\mathbf{x}}[\tilde{D}(G_{\text{dec}}(V(\mathbf{x})))]$, in (9) is equivalent to the third objective that forces $\tilde{D}$ to output a constant vector with equal value (i.e., probability) in each element.

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

1) Given the input image $\mathbf{x}$, D^{id} aims to correctly predict its identity label y^{id}.
2) Given the *real* prototype $\mathbf{x}^{\text{rp}}$ and the generated *fake* prototype by G, i.e., $\mathbf{x}^p = G_{\text{dec}}(P(\mathbf{x}))$, D^{gan} aims to classify $\mathbf{x}^{\text{rp}}$ as the real prototype and classify $\mathbf{x}^p$ as the fake prototype.

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

$$\max_D V_D = V_D^{\text{gan}} + \gamma V_D^{\text{id}} \tag{11}$$

where γ is a positive tradeoff parameter, and V_D^{id} and V_D^{gan} are defined as follows:

$$V_D^{\text{id}}(D^{\text{id}}, \mathbf{x}) = E_{\mathbf{x}}\left[\log D_{l^{\text{id}}}^{\text{id}}(\mathbf{x})\right] \tag{12}$$

$$V_D^{\text{gan}}(G, D^{\text{gan}}, \mathbf{x}^{\text{rp}}, \mathbf{x}) = E_{\mathbf{x}^{\text{rp}}}\left[\log D^{\text{gan}}(\mathbf{x}^{\text{rp}})\right] + E_{\mathbf{x}}\left[\log(1 - D^{\text{gan}}(G_{\text{dec}}(P(\mathbf{x}))))\right]. \tag{13}$$

For the discriminator $\tilde{D}$, the only purpose is to correctly predict the identity label for the generated variation, i.e., $\mathbf{x}^v = G_{\text{dec}}(V(\mathbf{x}))$. Formally, the objective function $V_{\tilde{D}}$ for training $\tilde{D}$ is as follows:

$$\max_{\tilde{D}} V_{\tilde{D}} = E_{\mathbf{x}}\left[\log \tilde{D}_{l^{\text{id}}}(G_{\text{dec}}(V(\mathbf{x})))\right] \tag{14}$$

where $\tilde{D}_i$ denotes the ith element in $\tilde{D}$.

Algorithm 1 DisP+V Training

Input: A training set of N_d identities with each image $\mathbf{x}$ annotated by the label $l = \{l^{\text{id}}, l^{\text{var}}\}$; A real prototype corpus with each image $\mathbf{x}^{\text{rp}}$ sampled from the distribution $\mathcal{P}_{\text{real}}$.

repeat
  Fix D and $\tilde{D}$, update G by solving the objective in Eq. (6)
  Fix G and $\tilde{D}$, update D by solving the objective in Eq. (11)
  Fix G and D, update $\tilde{D}$ by solving the objective in Eq. (14)
until convergence is achieved or a predefined maximum number of iterations is reached

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

For clarity, the training procedure of DisP+V is presented in **Algorithm 1**. It can be seen that we alternatively update the generator G, the discriminator D, and the discriminator $\tilde{D}$ by solving the objective functions V_G in (6), V_D in (11), and $V_{\tilde{D}}$ in (14) iteratively. During the alternative training process, G, D, and $\tilde{D}$ will be updated and improved. Specifically, with D^{gan} in D being more powerful in distinguishing real versus fake prototypes, G strives for generating a realistic-looking prototype in order to fool D^{gan}. Besides, D^{id} in D enables the generated prototype to preserve the identity characteristics and guides G_{enc} to learn a discriminative prototype feature that encodes as much identity information as possible. Furthermore, with $\tilde{D}$ being more powerful in classifying identity labels, G makes efforts to capture the less discriminative characteristics (i.e., facial variations) in $\mathbf{x}^v$ to fool $\tilde{D}$ to output a constant vector with a uniform distribution and guides G_{enc} to encode as little identity information as possible in the learned variation feature.

Generally speaking, there exist two adversarial learning processes between G, D, and $\tilde{D}$ in DisP+V. On the one hand, G and D compete with each other such that G disentangles a discriminative prototype feature relevant to identity in the latent space, and meanwhile, generating an identity-preserved prototype; on the other hand, G and $\tilde{D}$ also play an adversarial game which forces G to disentangle a less discriminative variation feature in the latent space and generating a variation image containing the corresponding facial variations. It is worth noting that, DisP+V introduces an extra discriminator $\tilde{D}$, while not directly using D^{id}, to predict the identity label for the generated variation $\mathbf{x}^v$. This strategy reduces the training complexity for D^{id} and enables D^{id} and $\tilde{D}$ to be responsible for their respective adversarial learning.

C. Application Scenarios

1) SSPP FR: Let $\mathbf{y}$ be a new query sample from the testing set, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]$ be the SSPP enrolment set with n identities, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_q]$ be the generic set with q samples from other identities *not of interest*. With the trained DisP+V model, we can obtain the prototype feature of $\mathbf{y}$, i.e., $P(\mathbf{y})$, and the prototype features of $\mathbf{S}$, i.e., $P(\mathbf{S})$. Subsequently, we classify the identity of $\mathbf{y}$ by matching $P(\mathbf{y})$ with $P(\mathbf{S}) = [P(\mathbf{s}_1), \ldots, P(\mathbf{s}_n)]$ as follows:

$$\text{Scheme 1: } \text{ID}(\mathbf{y}) = \arg\min_k \text{dist}(P(\mathbf{y}), P(\mathbf{s}_k)) \tag{15}$$

where $\text{dist}(\mathbf{a}, \mathbf{b})$ represents the distance between the feature vectors of $\mathbf{a}$ and $\mathbf{b}$, and the arccosine-distance, l_1-distance, or l_2-distance can be used as the distance metric.

Alternatively, we can also perform SSPP FR based on the P+V model in the latent space. With the trained DisP+V model, we further obtain the original feature of $\mathbf{y}$, i.e., $f(\mathbf{y}) = P(\mathbf{y}) + V(\mathbf{y})$, and the variation features of $\mathbf{A}$, i.e., $V(\mathbf{A})$. Then, we solve the following l_1-based optimization problem:

$$\begin{bmatrix} \boldsymbol{\alpha}^* \\ \boldsymbol{\beta}^* \end{bmatrix} = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\| f(\mathbf{y}) - [P(\mathbf{S}) \;\; V(\mathbf{A})] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_1 \tag{16}$$

where λ is a regularization parameter, $\boldsymbol{\alpha} \in \Re^n$ and $\boldsymbol{\beta} \in \Re^q$ are the coefficients of $P(\mathbf{S})$ and $V(\mathbf{A})$, respectively. In this article, (16) is solved via the basis pursuit denosing (BPDN)-homotopy algorithm [55]. Subsequently, $\mathbf{y}$ can be classified as the identity (i.e., class) according to the smallest reconstruction residual $r_k(\mathbf{y})$ among all classes

$$\text{Scheme 2: } \text{ID}(\mathbf{y}) = \arg\min_k r_k(\mathbf{y}) \tag{17}$$

where $r_k(\mathbf{y})$ is computed by

$$r_k(\mathbf{y}) = \left\| f(\mathbf{y}) - [P(\mathbf{S}) \;\; V(\mathbf{A})] \begin{bmatrix} \delta_k(\boldsymbol{\alpha}^*) \\ \boldsymbol{\beta}^* \end{bmatrix} \right\|_2^2 \tag{18}$$

with $\delta_k(\boldsymbol{\alpha}^*)$ being a vector whose nonzero entries are the entries in $\boldsymbol{\alpha}^*$ associated with class k.

To differentiate DisP+V using the two evaluation schemes, we denote DisP+V based on the latent-space P+V model in (16)–(18) as DisP+V_{pv} hereinafter. Furthermore, we analyze the time complexities of DisP+V and DisP+V_{pv} for recognizing the query sample $\mathbf{y}$, respectively. Specifically, both of the recognition stages for DisP+V and DisP+V_{pv} include two steps: feature extraction and classification. Suppose the image size of $\mathbf{y}$ is $w \times h$ and the number of the convolutional layers in G_{enc} is L. The time complexities of DisP+V and DisP+V_{pv} in feature extraction step are both $O(whL)$. In classification step, the time complexity of DisP+V is $O(dn)$, where d is the dimension of $P(\mathbf{y})$ and n is the size of the enrolment set $\mathbf{S}$, and the time complexity of DisP+V_{pv} is $O(\tau d^2 + \tau d(n+q))$ [8] where τ is the number of iterations for BPDN-homotopy in. (16) and q is the size of the generic set $\mathbf{A}$. Overall, the time complexities of DisP+V and DisP+V_{pv} in recognition stage are $O(whL+dn)$ and $O(whL+\tau d^2+\tau d(n+q))$, respectively. It is obvious that DisP+V costs less time than DisP+V_{pv} in recognition stage.

2) Other Applications: Besides the above-mentioned SSPP FR task, we can further leverage the trained generator G to do the following two tasks.

1) *Prototype Recovery:* Generating realistic-looking prototypes (e.g., an ID photograph) for contaminated samples in the enrolment database.
2) *Face Editing/Interpolation:* Performing semantic face editing/interpolation by modifying the disentangled variation feature in the latent space.

We will demonstrate the effectiveness of the proposed DisP+V regarding the above-mentioned potential applications, with extensive experiments and results in Section V.

V. Experimental Results

In this section, we start by detailing the experimental settings in Section V-A and then evaluate the proposed DisP+V by conducting the following experiments.

1) In Section V-B, we evaluate the recognition performance of DisP+V and DisP+V_{pv} for SSPP FR on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and Face Recognition Grand Challenge (FRGC) v2.0 datasets with four major single variations, i.e., expression, pose, disguise and lighting, and multiple variations.
2) In Section V-C, we evaluate the generated prototypes and the corresponding variation images by DisP+V on the above-mentioned five benchmark face datasets.
3) In Section V-D, we perform ablation study to investigate the roles of the D^{id}, D^{gan}, and $\tilde{D}$ on the performance of DisP+V.
4) In Section V-E, we evaluate the performance of our DisP+V for semantic face editing/interpolation.
5) In Section V-F, we further evaluate the performance of DisP+V when handling mixed facial variations on the unconstrained labeled faces in the wild (LFW)-a dataset. Moreover, we explore the feasibility of combining our DisP+V with the pretrained feature extractor for solving practical SSPP FR.

A. Experimental Settings

1) Dataset Description: Multi-PIE [56] is an extension of the Carnegie Mellon University Pose, Illumination, and Expression dataset [57] across multirecording sessions. It contains images of 337 identities under six different expressions across four sessions, 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.

FERET [34] is used for facial recognition system evaluation as part of the FERET program. It contains images of 1199 identities across ethnicity, gender, and age. We use 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.

CAS-PEAL [58] is constructed by the Joint R&D Laboratory for Advanced Computer and Communication Technologies, Chinese Academy of Sciences (CAS) Beijing, China. It contains 99 594 face images of 1040 identities with varying Pose, Expression, Accessory, and Lighting (PEAL). We use a subset of 300 identities from the Normal and Accessory categories, and thus, each identity has one neutral image and six images wearing different glasses and hats. We randomly choose 200 identities for training and the rest 100 ones for testing.

E-Yale-B is an extended version of the Yale Face Database B (Yale-B) [59]. It contains images of 38 identities under various lighting variations and is divided into five subsets. Subset 1, Subsets 2 and 3, and Subsets 4 and 5 characterize normal, slight-to-moderate, and severe lighting variations, respectively. AR [60] is created by Aleix Martinez and Robert Benavente, which contains images of 126 identities under variations of lighting, expression, and disguise. In the experiment, we merge E-Yale-B and AR lighting subset (100 identities) together to construct a new dataset, i.e., E-Yale-B&AR Light, to enrich the lighting variations. On this dataset, we randomly choose 100 identities for training and the rest 38 ones for testing.

FRGC v2.0 is the second version of the FRGC dataset [61], which contains 50 000 images of 4003 identities with two different facial expressions and under different illumination conditions. We use a subset of 150 identities with no less than 20 images per identity for evaluation. We randomly choose 100 identities for training and the rest 50 ones for testing.

LFW-a [62] is an aligned version of the LFW dataset [63] using a commercial face alignment software. It contains over 13 000 images of 5749 identities collected under uncontrolled environments with large variations in expressions, poses, illuminations, and so on. We use a subset of 158 identities with no less than ten images per identity for evaluation. We choose 50 identities containing neutral images for testing and use the rest 108 ones for training.

TABLE II
Network Structure of G

G_{enc}		
Layer	Filter / Stride / Pad	Output Size
Conv1	3 × 3 / 1 / 1	96×96×32
Conv2	3 × 3 / 1 / 1	96×96×64
Conv3	3 × 3 / 2 / 0	48×48×64
Conv4	3 × 3 / 1 / 1	48×48×64
Conv5	3 × 3 / 1 / 1	48×48×128
Conv6	3 × 3 / 2 / 0	24×24×128
Conv7	3 × 3 / 1 / 1	24×24×96
Conv8	3 × 3 / 1 / 1	24×24×192
Conv9	3 × 3 / 2 / 0	12×12×192
Conv10	3 × 3 / 1 / 1	12×12×128
Conv11	3 × 3 / 1 / 1	12×12×256
Conv12	3 × 3 / 2 / 0	6×6×256
Conv13	3 × 3 / 1 / 1	6×6×160
Conv14	3 × 3 / 1 / 1	6×6×320
Conv15-1, Conv15-2	3 × 3 / 1 / 1	6×6×320
Conv16-1, Conv16-2	3 × 3 / 1 / 1	6×6×320
Conv17-1, Conv17-2	3 × 3 / 1 / 1	6×6×256
Pool18-1, Pool18-2	6 × 6 / 1 / 0	1×1×256
G_{dec}		
FC-19	–	6×6×256
DeConv20	3 × 3 / 1 / 1	6×6×160
DeConv21	3 × 3 / 1 / 1	6×6×256
DeConv22	3 × 3 / 2 / 0	12×12×256
DeConv23	3 × 3 / 1 / 1	12×12×128
DeConv24	3 × 3 / 1 / 1	12×12×192
DeConv25	3 × 3 / 2 / 0	24×24×192
DeConv26	3 × 3 / 1 / 1	24×24×96
DeConv27	3 × 3 / 1 / 1	24×24×128
DeConv28	3 × 3 / 2 / 0	48×48×128
DeConv29	3 × 3 / 1 / 1	48×48×64
DeConv30	3 × 3 / 1 / 1	48×48×64
DeConv31	3 × 3 / 2 / 0	96×96×64
DeConv32	3 × 3 / 1 / 1	96×96×32
DeConv33	3 × 3 / 1 / 1	96×96×3

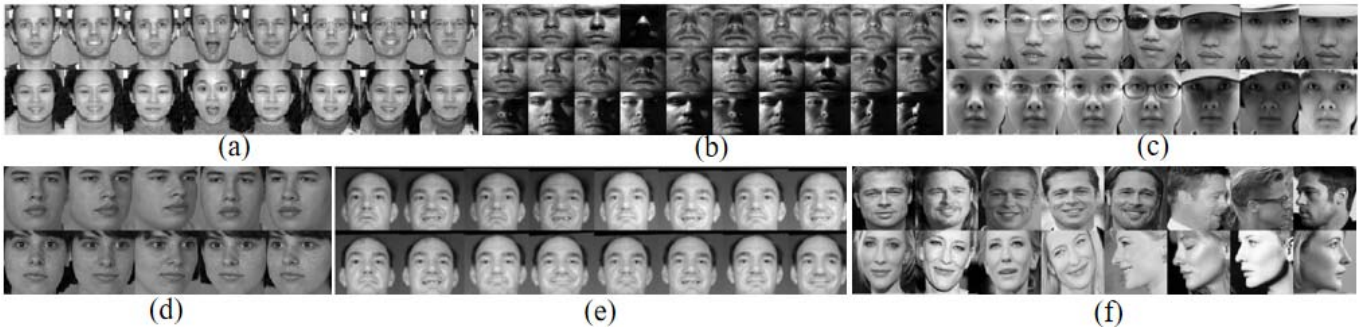


Fig. 3. Illustration of some gray face examples from six constrained and unconstrained datasets: (a) Multi-PIE. (b) E-Yale-B. (c) CAS-PEAL. (d) FERET. (e) FRGC v2.0. (f) LFW-a.

TABLE III
NETWORK STRUCTURES OF D AND $\tilde{D}$

D and $\tilde{D}$		
Layer	Filter / Stride / Pad	Output Size
Conv1	3 × 3 / 1 / 1	96×96×32
Conv2	3 × 3 / 1 / 1	96×96×64
Conv3	3 × 3 / 2 / 0	48×48×64
Conv4	3 × 3 / 1 / 1	48×48×64
Conv5	3 × 3 / 1 / 1	48×48×128
Conv6	3 × 3 / 2 / 0	24×24×128
Conv7	3 × 3 / 1 / 1	24×24×96
Conv8	3 × 3 / 1 / 1	24×24×192
Conv9	3 × 3 / 2 / 0	12×12×192
Conv10	3 × 3 / 1 / 1	12×12×128
Conv11	3 × 3 / 1 / 1	12×12×256
Conv12	3 × 3 / 2 / 0	6×6×256
Conv13	3 × 3 / 1 / 1	6×6×160
Conv14	3 × 3 / 1 / 1	6×6×320
Pool15	6 × 6 / 1 / 0	1×1×320
FC16 (D)	–	N_d+1
FC16 ($\tilde{D}$)	–	N_d

For each dataset, all face samples are first aligned to a canonical view of size 100 × 100 and then center cropped to 96 × 96. We show some gray face samples on Multi-PIE, E-Yale-B, CAS-PEAL, FERET, FRGC v2.0, and LFW-a face datasets in Fig. 3.

2) Implementation Details: In the first, we introduce the network structures of the generator G and the two discriminators D and $\tilde{D}$.

For the generator G, we adopt the CASIA-Net in [52] as the backbone of G_{enc} and G_{dec}, where batch normalization and exponential linear unit are used after each `conv` and `deconv` layer. In G_{enc}, the final `AvgPool` layer is replaced by two subnets with each having three `conv` layers and one global `AvgPool`. The two disentanglement branches extract two 256-D features for $\mathbf{x}$, i.e., $P(\mathbf{x})$ and $V(\mathbf{x})$. Subsequently, $P(\mathbf{x})$, $V(\mathbf{x})$, and their superposition, i.e., $f(\mathbf{x}) = P(\mathbf{x}) + V(\mathbf{x})$, are used as the inputs for G_{dec} to generate the prototype $\mathbf{x}^p$, the variation image $\mathbf{x}^v$, and the reconstructed image $\widehat{\mathbf{x}}$ for $\mathbf{x}$, respectively. The network structure of G is presented in Table II.

For the discriminators D and $\tilde{D}$, they both have an extra fully connection (`FC`) layer based on CASIA-Net. The output of D is a $(N_d + 1)$-dimensional vector, where the first N_d elements are the outputs of D^{id} for predicting the face identity and the rest one is reserved for D^{gan} to distinguish real versus fake prototype. The output of $\tilde{D}$ is a N_d-dimensional vector only for face identity prediction. The network structures of D and $\tilde{D}$ are presented in Table III.

TABLE IV
DATASET PARTITION AND PARAMETER SETTING

Dataset	#Train. identity	#Test. identity	N_d	Trade-off parameter
Multi-PIE	100	41	100	μ_1=γ=5.0 μ_2=0.5 μ_3=0.1
FERET	150	50	150	
CAS-PEAL	200	100	200	
E-Yale-B&AR Light	100	38	100	
FRGC v2.0	100	50	100	
LFW-a	108	50	108	

We train the proposed DisP+V[4] by the mini-batch stochastic gradient descent with a mini-batch size of 16. The maximum number of training epochs is set as 2000. All weights are initialized from a zero-centered normal distribution with the standard deviation of 0.02. Following the work in [52], we adopt the Adam optimizer [64] with tuned hyperparameters for optimizing, where the learning rate and momentum are empirically set as 0.0002 and 0.5, respectively.

3) Parameter Setting: For each evaluated dataset, N_d is set as the total number of identities in the training set. We tune all tradeoff hyperparameters via grid search. Specifically, we observe that DisP+V achieve promising performance when the tradeoff parameters μ_1, μ_2, and μ_3 in (6) and γ in (11) are set at 5.0, 0.5, 0.1, and 5.0, respectively, and fix the values across all datasets. Moreover, the number of training and testing identities in each dataset are also specified. All the above-mentioned parameter settings and training/testing sets partition are detailed in Table IV.

B. Evaluation on SSPP FR

This section evaluates the recognition performance of DisP+V and DisP+V_{pv} for SSPP FR (including SSPP-ce FR and SSPP-se FR) on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 datasets. For DisP+V and DisP+V_{pv}, we adopt the evaluation schemes in (15) and (17), respectively, to perform SSPP FR.

On each dataset, we randomly choose one sample (could be a standard sample or a contaminated sample) for each identity to construct the contaminated enrolment database in SSPP-ce FR and use the rest as the query samples for recognition. We set the contaminated ratio (i.e., #contaminated samples/#total identities) ranging from 10% to 90% with an

[4]The code is released at https://github.com/PangMeng92/DisPV_Code.git.

TABLE V

RANK-1 RECOGNITION RATES (%) ± STANDARD ERRORS (%) AND THE RECOGNITION TIME (S) OF DIFFERENT METHODS ON THE MULTI-PIE, FERET, CAS-PEAL, E-YALE-B&AR LIGHT, AND FRGC V2.0 DATASETS FOR SSPP FR. IN THE BRACKETS, WE SHOW THE IMPROVEMENT OF OUR DISP+V_{PV} AND DISP+V OVER THE SECOND BEST METHOD IN THE CASE. * INDICATES STATISTICAL SIGNIFICANCE WITH p-VALUE < 0.05

Enrolment database		SRC	VAE	SVDL	SLRC	S^3RC	DisP+V_{pv}	DisP+V
Multi-PIE [Expression]	Ratio=0.0	67.8	59.0	78.4	75.9	79.8	83.2* (↑ 3.4)	83.4* (↑ 3.6)
	Ratio=0.1	66.7±2.6	57.0±1.5	77.3±1.4	75.2±2.4	77.9±2.1	83.0*±0.7 (↑ 5.1)	83.1*±1.4 (↑ 5.2)
	Ratio=0.3	64.7±1.7	56.4±0.7	74.5±3.2	74.8±2.2	76.6±0.9	80.7*±1.9 (↑ 4.1)	81.3*±1.9 (↑ 4.7)
	Ratio=0.5	66.3±2.2	56.5±1.4	73.1±1.1	73.9±0.9	75.5±1.9	79.5*±0.8 (↑ 4.0)	79.8*±1.7 (↑ 4.3)
	Ratio=0.7	64.4±2.8	56.5±2.5	70.8±1.7	70.9±3.4	73.3±4.7	79.8*±3.5 (↑ 6.5)	79.8*±4.4 (↑ 6.5)
	Ratio=0.9	63.7±2.7	56.3±2.4	69.2±3.1	68.8±3.4	71.6±4.7	78.7*±3.4 (↑ 7.1)	79.5*±2.6 (↑ 7.9)
	Time	0.0032	0.0169	0.0132	0.0380	0.0930	0.0452	0.0333
FERET [Pose]	Ratio=0.0	51.5	55.0	67.0	68.0	73.0	91.0* (↑ 18.0)	93.5* (↑ 20.5)
	Ratio=0.1	37.2±2.6	47.9±2.3	61.3±1.7	62.7±3.8	73.9±2.1	91.0*±0.7 (↑ 17.1)	92.8*±1.2 (↑ 18.9)
	Ratio=0.3	30.4±2.0	37.4±3.1	53.5±1.7	59.1±3.2	70.3±3.9	90.2*±2.1 (↑ 19.9)	92.8*±1.6 (↑ 22.5)
	Ratio=0.5	30.4±1.5	33.4±1.6	50.5±3.5	56.8±3.0	66.9±3.0	89.3*±1.4 (↑ 22.4)	92.0*±1.6 (↑ 25.1)
	Ratio=0.7	29.5±0.9	32.2±1.2	47.0±1.7	54.5±2.7	63.6±1.2	88.9*±1.6 (↑ 25.3)	91.2*±2.1 (↑ 27.6)
	Ratio=0.9	28.9±1.9	29.1±0.9	45.6±1.8	50.9±2.4	64.8±3.4	89.0*±1.9 (↑ 24.2)	90.8*±2.0 (↑ 26.0)
	Time	0.0042	0.0186	0.0124	0.0412	0.0934	0.0445	0.0392
CAS-PEAL [Disguise]	Ratio=0.0	62.3	51.4	78.7	78.2	80.3	86.0* (↑ 5.7)	84.7* (↑ 4.4)
	Ratio=0.1	54.7±1.4	45.3±0.9	76.7±1.2	76.5±1.3	77.0±3.3	83.5*±0.9 (↑ 6.5)	83.0*±1.2 (↑ 6.0)
	Ratio=0.3	51.4±0.8	39.9±1.5	72.1±1.9	70.0±1.2	72.4±2.7	80.1*±1.8 (↑ 7.7)	79.7*±1.6 (↑ 7.3)
	Ratio=0.5	47.1±2.1	35.2±2.0	67.9±2.1	65.9±1.2	67.6±2.3	77.1*±0.4 (↑ 9.2)	75.6*±1.3 (↑ 7.7)
	Ratio=0.7	38.3±2.6	28.8±1.6	60.3±2.9	59.1±3.2	61.8±2.6	70.2*±1.3 (↑ 8.4)	69.5*±3.6 (↑ 7.7)
	Ratio=0.9	39.7±2.5	29.1±2.0	59.6±2.4	57.3±1.1	60.0±3.5	70.1*±2.2 (↑ 10.1)	69.4*±2.3 (↑ 9.4)
	Time	0.0057	0.0153	0.0143	0.0451	0.1075	0.0476	0.0376
E-Yale-B&AR Light [Lighting]	Ratio=0.0	64.0	59.9	88.1	88.8	88.2	91.4* (↑ 2.6)	92.3* (↑ 3.5)
	Ratio=0.1	51.6±3.8	53.9±1.1	84.3±1.8	87.3±2.6	81.2±1.5	90.0*±1.8 (↑ 2.7)	90.9*±1.3 (↑ 3.6)
	Ratio=0.3	48.3±4.8	50.5±2.1	81.3±3.4	83.4±2.4	79.5±5.0	88.2*±2.9 (↑ 4.8)	88.4*±2.5 (↑ 5.0)
	Ratio=0.5	49.6±3.6	47.3±2.4	80.4±4.2	81.4±4.1	75.3±2.9	88.6*±1.7 (↑ 7.2)	88.9*±1.7 (↑ 7.5)
	Ratio=0.7	47.2±2.2	40.1±1.7	75.5±2.6	77.8±2.2	71.8±3.9	86.6*±2.9 (↑ 8.8)	87.4*±3.4 (↑ 9.6)
	Ratio=0.9	45.3±4.0	38.8±3.8	73.8±5.7	74.6±4.6	64.6±4.7	85.2*±3.1 (↑ 10.6)	85.0*±2.5 (↑ 10.4)
	Time	0.0039	0.0137	0.0114	0.0411	0.0958	0.0416	0.0326
FRGC v2.0 [Multiple variations]	Ratio=0.0	72.4	52.4	85.3	85.0	86.5	89.4* (↑ 2.9)	88.2* (↑ 1.7)
	Ratio=0.1	71.5±0.5	50.5±2.2	83.7±0.6	84.6±1.3	85.9±1.5	88.2*±0.8 (↑ 2.3)	86.7*±1.0 (↑ 0.8)
	Ratio=0.3	70.6±1.7	49.8±1.4	83.2±1.2	82.7±1.3	85.1±1.0	87.7±2.4 (↑ 2.6)	85.9±1.3 (↑ 0.8)
	Ratio=0.5	70.3±2.5	51.1±2.3	82.7±1.5	82.1±2.1	83.8±1.7	87.2*±2.3 (↑ 3.4)	85.8*±2.3 (↑ 2.0)
	Ratio=0.7	71.3±1.4	53.9±3.6	83.0±1.2	82.1±1.1	83.9±1.1	87.4*±1.5 (↑ 3.5)	85.6*±1.2 (↑ 1.7)
	Ratio=0.9	71.7±1.1	53.9±1.3	81.5±1.2	80.7±1.3	82.7±1.1	86.7*±1.1 (↑ 4.0)	85.9*±1.5 (↑ 3.2)
	Time	0.0026	0.0119	0.0106	0.0245	0.0889	0.0399	0.0288

interval of 20%. We repeat each experiment five times and report the average results. Furthermore, we also present the recognition results when the contaminated ratio is zero, which is exactly the setting of SSPP-se FR.

We choose five representative methods for comparison, including the baseline SRC [22], the representation learning-based VAE [48], two recent generic learning methods, i.e., SLRC [26] and SVDL [28], and the latest prototype learning S^3RC [43] method. For SVDL, SLRC, and S^3RC, the training set is used as the auxiliary generic set for generating variation dictionaries. We tune the regularization parameter λ of SRC, SLRC, and S^3RC and find that they achieve the best performance when $\lambda = 0.01$. For SVDL, as suggested in [28], the parameters λ_1, λ_2, and λ_3 are set at 0.001, 0.01, and 0.0001, respectively. For DisP+V_{pv}, the regularization parameter λ in (16) is set at 0.1. For VAE and DisP+V, the arccosine-distance metric is used for measuring the distance between two representations.

Table V lists the rank-1 recognition rates (±standard errors) and the recognition time of all the methods on the five datasets for SSPP FR. Furthermore, we report the statistical significance between the recognition results of our proposed methods (including DisP+V_{pv} and DisP+V) and that of the second-best method in each case by comparing their p-values [65] with the significance level of 0.05. From Table V, we have the following key observations.

1) Our proposed DisP+V_{pv} and DisP+V consistently obtain higher rank-1 recognition rates than the other compared methods for both SSPP-se FR (ratio = 0%) and SSPP-ce FR (ratio > 0%) in all cases across the five datasets. Moreover, the improvements of our proposed methods over the second-best method in each case are statistically significant as the corresponding p-values < 0.05.
2) As the enrolment contamination ratio rises from 0% to 90%, more enrolment samples are contaminated and incorrectly represent the personal identities. Under the circumstances, the recognition accuracies of all the methods tend to decrease. However, DisP+V_{pv} and DisP+V have shown greater robustness against the enrolment contamination increase than the other compared methods, and the advantages become more obvious when the ratio reaches higher. The superiority of our DisP+V and DisP+V_{pv} attributes to the successful disentanglement of the prototype and variation features in the latent space, which enables the learned

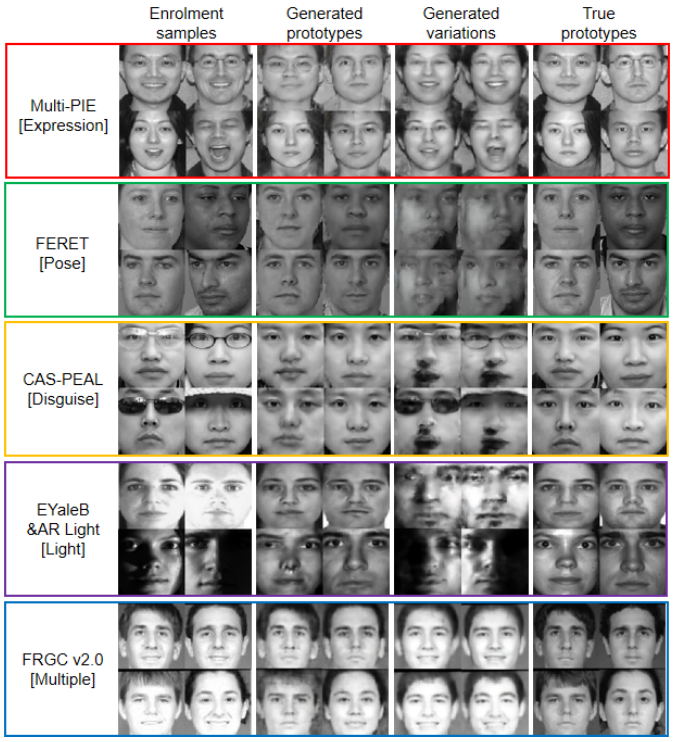


Fig. 4. Generated prototypes and variations of some selected examples by DisP+V on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 datasets. Figures from left to right are: original enrolment samples, generated prototypes, generated variations, and true prototypes.

prototype feature to encode as much identity information as possible.

3) It is interesting to find that each of DisP+V (based on direct prototype feature matching) and DisP+V_{pv} (based on latent-space P+V model recognition) has its own advantage when handling different facial variations. For example, DisP+V performs better on FERET with pose variations, while DisP+V_{pv} is better at handling additive variations, such as disguise on CAS-PEAL.
4) S^3RC usually performs better than the generic learning SVDL and SLRC methods with the contamination because it involves a prototype learning step for restoring contaminated enrolment samples. However, S^3RC obtains poor performance and is inferior to SVDL and SLRC on E-Yale-B&AR Light. The reason is that the quality of the learned prototypes by S^3RC depends heavily on the clustering performance of GMM, which is sensitive to severe lightings and shadows.
5) SVDL and SLRC obtain close results as they both use the classic P+V model for recognition. They perform poorly on FERET because the used P+V model is a linear superposition model in the pixel-spatial space, which is less effective in handling the nonlinear pose variation. In contrast, our proposed methods achieve much better recognition results. For example, DisP+V delivers 25.5%, 30.1%, 33.7%, 35.2%, 36.7%, and 39.9% improvements over SLRC when the enrolment contaminated ratio is set at 0%, 10%, 30%, 50%, 70%, and 90%, respectively.
6) Although the representation learning-based VAE also performs variation disentanglement during encoding, it is much less competitive with our DisP+V and DisP+V_{pv}. This is because it is an unsupervised method and does not exploit the labeled identity information.
7) The recognition time of DisP+V on each dataset is less than that of DisP+V_{pv}, which is consistent with the complexity analysis results in Section IV-C. Moreover, the recognition time of DisP+V and DisP+V_{pv} are both far less than the acceptable 0.5 s, which is applicable from a real-time perspective. VAE costs less time than DisP+V as its encoder (i.e., feature extractor) has fewer convolutional layers. S^3RC costs more time than SLRC and SVDL because it has an extra GMM clustering process for prototype learning. In addition, SRC costs the least time among all the methods.

TABLE VI

VERIFICATION PERFORMANCE OF DISP+V ON THE MULTI-PIE, FERET, CAS-PEAL, E-YALE-B&AR LIGHT, AND FRGC V2.0 DATASETS

Dataset	TPR(%)@FAR=0.1		AP (%)	
	Baseline	DisP+V	Baseline	DisP+V
Multi-PIE [Expression]	60.8±1.0	**61.0±2.1**	77.4±1.1	**79.3±1.1**
FERET [Pose]	44.6±3.1	**52.4±2.2**	66.9±2.2	**70.4±0.9**
CAS-PEAL [Disguise]	49.7±3.2	**56.7±3.2**	66.3±1.9	**76.6±2.4**
E-Yale-B&AR Light [Lighting]	63.9±2.2	**80.0±0.8**	75.6±0.9	**90.0±1.3**
FRGC v2.0 [Multiple variations]	78.0±2.5	**83.2±1.7**	89.7±0.7	**89.9±0.7**

C. Evaluation on Generated Prototype and Variation

This section evaluates the generated prototypes and the variations by our proposed DisP+V on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 datasets. In the experiments, the quality of the generated prototypes is measured from both qualitative and quantitative perspectives.

1) Qualitative Analysis Results: We first illustrate the generated prototypes and the variations for four random enrolment samples on each dataset in Fig. 4. For reference, we also show the true prototypes of these enrolment samples.

From Fig. 4, we can observe that the prototypes and the corresponding variations are well disentangled from the contaminated enrolment samples on all five datasets. Intuitively, for enrolment samples contaminated by a single variation, such as expression, pose, disguise, or lighting, DisP+V successfully removes the corresponding variation in the learned prototypes. Even in the case where the enrolment sample on FRGC v2.0 is contaminated by multiple variations and the input type of variation is unknown in advance, our DisP+V can still recover appropriate prototypes to represent the identities. Furthermore, we also observe that the generated variations capture the facial variations of the original enrolment samples correctly and contain little input identity information.

2) Quantitative Analysis Results: Since most of the generated prototypes by our DisP+V are visually appealing, it is expected that these learned prototypes are more suitable to represent the identities than the original contaminated enrolment samples. To verify this assumption, we further perform verification experiments between the learned prototypes by DisP+V and the true prototypes and compare them with the verification results between the original enrolment samples and the true prototypes (baseline). Specifically, for each dataset, we randomly sample 600 pairs of the generated prototypes and

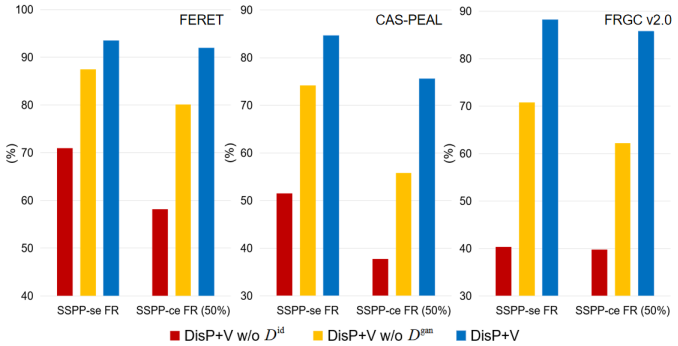


Fig. 5. Comparison results of DisP+V and its variants DisP+V w/o D^{id} and DisP+V w/o D^{var} on FERET, CAS-PEAL, and FRGC v2.0 datasets.

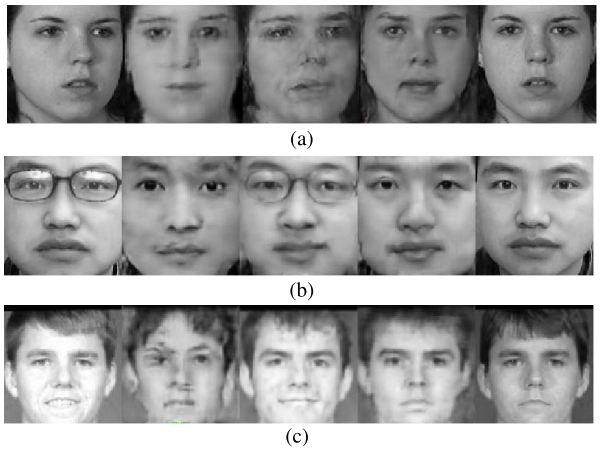


Fig. 6. Prototype learning examples of DisP+V and its two variants on (a) FERET, (b) CAS-PEAL, and (c) FRGC v2.0 datasets. The figures from left to right are the original enrolment sample, the generated prototype by DisP+V w/o D^{id}, the generated prototype by DisP+V w/o D^{gan}, the generated prototype by DisP+V, and the true prototype for reference, respectively.

true prototypes, where 200 pairs are positive and the remaining 400 pairs are negative, for verification. The cosine similarity between each pair of samples is used for verification.

Two common-used metrics, i.e., true positive rate (TPR) and average precision, are employed to measure the verification performance. For the detailed definitions of the two metrics, please refer to [66]–[68]. For TPR, we tune the similarity threshold to let the false acceptance rate be 0.1. Each verification experiment is repeated five times and the average results ($\pm$ standard errors) on the five evaluated datasets are presented in Table VI. It can be observed that our DisP+V consistently achieves better verification performance than the baseline method in all cases over the five evaluated datasets, which indicates that: 1) the generated prototypes by our DisP+V preserve the input identity characteristics well and 2) are closer to the true prototypes than the original contaminated enrolment samples.

D. Ablation Study

In this section, we perform an ablation study on DisP+V. In DisP+V, there are two discriminators, i.e., $D = [D^{\text{id}}, D^{\text{gan}}]$ and $\tilde{D}$. We first investigate the roles of D^{id} and D^{gan} in D on the performance of DisP+V. Accordingly, we construct two variants of DisP+V by removing D^{id} and D^{gan} and denote them as DisP+V w/o D^{id} and DisP+V w/o D^{gan}, respectively. We compare DisP+V with DisP+V w/o D^{id} and DisP+V w/o D^{gan} in terms of the recognition accuracy on the FERET, CAS-PEAL, and FRGC v2.0 datasets that contain pose, disguise, and multiple variations of expression and lighting.

[5]More face editing and interpolation results are available at https://github.com/PangMeng92/DisPV_TNNLS_Supplementary.git.

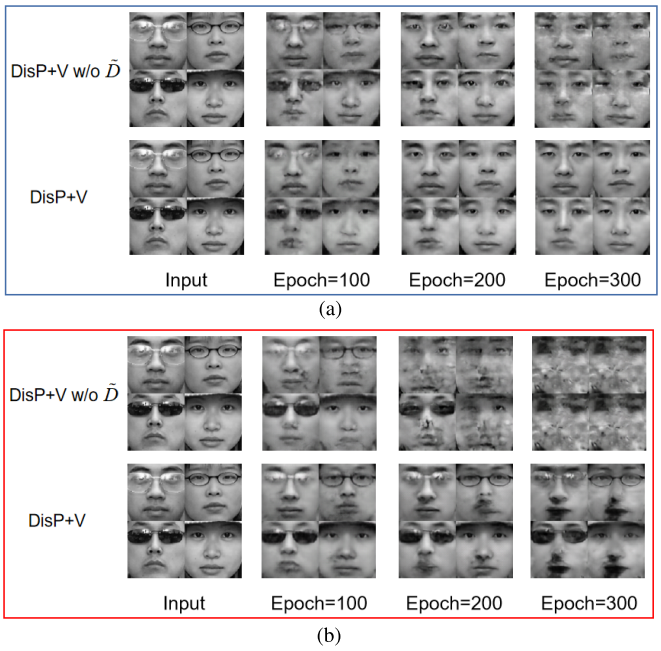


Fig. 7. Examples of generated prototypes and variations by DisP+V and DisP+V w/o $\tilde{D}$ on CAS-PEAL dataset when the number of training epoches increases from 100 to 300. (a) Generated prototypes. (b) Generated variations.

As shown in Fig. 5, DisP+V consistently outperforms the two variants over the three datasets. For example, DisP+V delivers 33.8% (or 11.9%), 37.9% (or 19.8%), and 45.9% (or 23.6%) improvements over DisP+V w/o D^{id} (or DisP+V w/o D^{gan}) on FERET, CAS-PEAL and FRGC v2.0, respectively, w.r.t. recognition rate for SSPP-ce FR with the contaminated ratio of 50%. The results show that both of D^{id} and D^{gan} contribute to the recognition performance of DisP+V. Moreover, we observe that D^{id} plays a more important role than D^{gan} as DisP+V w/o D^{id} suffers larger performance degradation. This is because D^{id} is used to preserve the identity label, which captures the most important identity information. Furthermore, we illustrate the generated prototypes of an example input image by DisP+V and the two variants on FERET, CAS-PEAL, and FRGC v2.0, respectively, in Fig. 6. We can see that, when removing D^{id}, the identity characteristics of the input sample are not preserved well in the generated prototype or even difficult to be recognized; when removing D^{gan}, the variation still exists in the generated prototype.

Subsequently, we study the role of $\tilde{D}$. As mentioned earlier, we introduce this extra $\tilde{D}$ for predicting the ID of the variation $\mathbf{x}^v$ individually. In the experiment, we remove $\tilde{D}$ and directly use D^{id} to predict both IDs of $\mathbf{x}^p$ and $\mathbf{x}^v$ and perform two adversarial learning based on D^{id}. We denote the DisP+V variant as DisP+V w/o $\tilde{D}$ and illustrate the generated prototypes and variations by DisP+V and DisP+V w/o $\tilde{D}$ on CAS-PEAL when the number of training epoch equals 100, 200, and 300, respectively, in Fig. 7. It can be observed that: 1) compared with DisP+V w/o $\tilde{D}$, DisP+V usually generates visually better prototypes containing fewer artifacts and more accurate facial variations and 2) DisP+V w/o

Fig. 8. Face editing and interpolation results[5] on (a) Multi-PIE (Expression), (b) FERET (Pose), (c) CAS-PEAL (Disguise), (d) E-Yale-B&AR Light (Lighting), and (e) FRGC v2.0 (Multiple variations). The two figures in the leftmost column are the input face images (one is standard and the other contains variation). The figures from top to bottom are face images of other identities containing different target variations, the edited images, and the interpolated images, respectively.

TABLE VII
RANK-10 RECOGNITION RATES OF OUR DISP+V AND THE COMPARED GENERIC LEARNING AND PROTOTYPE LEARNING METHODS FOR BOTH SSPP-SE FR AND SSPP-CE FR (RATIO = 50%) ON LFW-A DATASET

Method	SSPP-se FR	SSPP-ce FR
SRC	50.2	47.3±1.5
SVDL	57.6	52.9±1.6
SLRC	65.1	62.6±2.5
S^3RC	65.6	62.9±2.3
DisP+V	**72.2**	**70.9±1.6**

$\tilde{D}$ generates visually terrible prototypes and noninformative variations when the number of training epochs reaches 300, which indicates that one single discriminator D^{id} is insufficient to tolerate two different adversarial training in DisP+V and verifies the rationality of introducing the extra $\tilde{D}$ for dealing with $\mathbf{x}^v$.

E. Face Editing/Interpolation

In this section, we explore the feasibility of DisP+V for semantic face editing and interpolation [69]–[71]. To this end, we take several face images on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 datasets and edit (or interpolate) them by replacing their disentangled variation features with the ones extracted from the target identities.

Fig. 8(a)–(e) shows some examples of face editing and interpolation results on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 datasets, respectively. From Fig. 8(a)–(e), we have two key observations.

1) DisP+V demonstrates powerful face editing ability on adding target facial variations such as different expressions (e.g., smile, laugh, and disgust), different poses, different disguises (e.g., glasses and hat), different lightings, or multiple variations of expressions and lightings into the standard images with little artifacts.
2) DisP+V is also capable of interpolating face images by changing the original variations into the target ones. For instance, in Fig. 8(c), the light-color ordinary glasses in the input face have been well replaced by the other types of ordinary glasses, sunglasses, and hats successively, and the corresponding interpolated images look natural.

F. Evaluation Under Unconstrained Environment

In practice, an enrolment sample is likely to be contaminated by complex mixed variations such as the combination of two or more different variations. In this section, we apply our DisP+V to the unconstrained LFW-a dataset that contains various mixed variations in the wild and evaluates its recognition performance for SSPP FR in an unconstrained setting.

We first compare DisP+V with the baseline SRC, the generic learning SLRC and SVDL, and the prototype learning S^3RC. The parameters of DisP+V and the other methods are set in the same way as in Section V-B. We list the rank-10 recognition rates of all the methods for SSPP-se FR and SSPP-ce FR (ratio = 50%) in Table VII. Furthermore, we enhance DisP+V by replacing the original encoder with a pretrained LightCNN-29 feature extractor [72] on CASIA-WebFace [73] and MS-Celeb-1M [74] datasets. We enforce the dimension of the extracted features still to be 256 by modifying the two disentanglement branches as two three-layer FC (input: 256, output: 256) nets. The network structures of the decoder G_{dec} in G and the two discriminators D and $\tilde{D}$ are kept unchanged. In training, we freeze the parameters' values in the LightCNN-29 but just update the parameters' values of the FC layers, G_{dec}, D, and $\tilde{D}$. We denote our DisP+V using the LightCNN-29 feature extractor as DisP+V_{LC29}, and add five recent deep learning-based methods, i.e., DeepID [75], joint and collaborative representation with local adaptive convolution feature (JCR-ACF) [5], VGG-face [76], regularface [37], Arc-face [36], and the state-of-the-art class-level joint representation with regional adaptive convolution feature (CJR-RACF) [11], for comparison. We follow the evaluation protocol in JCR-ACF and present the rank-1 recognition rates of DisP+V_{LC29} and the compared deep learning-based methods in Table VII. From Tables VII and VIII, we have the following key observations.

1) There exists a large gap between the performance of DisP+V, SLRC, SVDL, and S^3RC in Table VII and that

TABLE VIII

RANK-1 RECOGNITION RATES (%) OF OUR DISP+V_{LC29} AND THE COMPARED DEEP LEARNING-BASED METHODS ON LFW-A DATASET

Methods	Accuracy (%)
DeepID	70.7
JCR-ACF	86.0
VGG-face	84.7
Regular-face	83.7
Arc-face	92.3
CJR-RACF	95.5
DisP+V_{LC29}	**96.7**

Fig. 9. Prototype learning examples of nine selected enrolment samples on LFW-a. From top to bottom: (a) original enrolment samples, (b) our generated prototypes, and (c) true prototypes for reference.

in Table V, which indicates that it is rather challenging to perform SSPP FR with mixed variations based on a small-scale partitioned training set. In this case, our DisP+V still outperforms the compared generic learning SLRC and SVDL, and the prototype learning S^3RC.

2) By introducing related large-scale Web face datasets as the auxiliary set for pretraining, the five deep learning-based methods obtain promising results based on the pretrained models/features. Particularly for CJR-RACF, it achieves a high rank-1 recognition rate of 95.5%.
3) Benefiting from the pretrained LightCNN-29 feature extractor, DisP+V_{LC29} has a significant gain over DisP+V and achieves an inspiring recognition rate of 96.7% for SSPP FR on LFW-a, which is better than 95.5% obtained by the state-of-the-art CJR-RACF.

Furthermore, we visualize the colored generated prototypes by DisP+V for nine contaminated enrolment samples in Fig. 9. It can be observed that our DisP+V shows good capabilities to learn identity-preserved prototypes for the samples with the mixed variations of slight-to-moderate poses and expressions. It is worth mentioning that in a few cases where enrolment samples are contaminated by serious facial variations, such as mixed variations of large poses and expressions/occlusions, DisP+V cannot generate satisfactory prototypes because some key facial information is missing in these cases.

Generally speaking, the experimental results in Fig. 9 and Table VII have demonstrated the effectiveness of DisP+V to learn prototypes for the in-the-wild faces containing complex mixed variations and the superiority for performing SSPP FR in unconstrained setting over the existing generic learning and prototype learning methods. Moreover, the significant improvement of DisP+v_{LC29} over DisP+V verifies the feasibility of combining our DisP+V with pretrained deep feature extractors for solving practical SSPP FR.

VI. CONCLUSION

In this article, we have proposed a new *disentangled* prototype plus variation (DisP+V) model. In contrast to the classic P+V model that combines face images in the observational pixel-spatial space and can only handle linear variations, our DisP+V performs the combination in a latent semantic space and can handle both linear and nonlinear variations. DisP+V consists of an encoder–decoder structural generator and two discriminators. The generator and discriminators play two adversarial games such that the generator: 1) nonlinearly encodes the images into a latent semantic space where the more discriminative prototype feature and the less discriminative variation feature are disentangled and 2) generating an identity-preserved prototype and the corresponding variation image. Extensive experiments on various real-world face datasets with single/multiple and mixed variations have verified the superiority of DisP+V over the classic P+V model-based counterparts for SSPP FR and the effectiveness for handling the tasks of prototype recovery and face editing/interpolation.

It is worth mentioning that, although the proposed DisP+V has shown the promising ability for learning homogeneous prototype from a contaminated face image in a single domain, it is unable to learn heterogeneous prototypes across different domains (e.g., near infrared→visible) because it ignores considering the key factor of domain type. Such a new issue of heterogeneous prototype learning (HPL) is quite challenging as it involves two intertwined subproblems of prototype learning and domain transfer. To tackle HPL, we aim to generalize DisP+V to multiple domains based on a new face composition hypothesis (i.e., P+V+D model) that a face image is composed by the three factors of identity-relevant prototype, facial variation, and domain type. We will leave the interesting study as the future research work.

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] H. Deng, W. Chen, Q. Shen, A. J. Ma, P. C. Yuen, and G. Feng, "Invariant subspace learning for time series data based on dynamic time warping distance," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107210.
[8] M. Pang, Y.-M. Cheung, Q. Shi, and M. Li, "Iterative dynamic generic learning for face recognition from a contaminated single-sample per person," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1560–1574, Apr. 2021.
[9] 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, Jan. 2020, doi: 10.1109/TIFS.2020.2970590.
[10] 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, Jun. 2020, doi: 10.1109/TIFS.2020.3001665.

[11] 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, Jan. 2020, doi: 10.1109/TIFS.2020.2965301.

[12] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1445–1451, Apr. 2021.

[13] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3D room layout estimation from a single RGB image," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 3014–3024, Nov. 2020.

[14] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, and Y. Zhang, "Depth image denoising using nuclear norm and learning graph model," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 4, pp. 1–17, Jan. 2021.

[15] 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 Trans. Inf. Forensics Security*, vol. 16, pp. 2246–2259, Jan. 2021, doi: 10.1109/TIFS.2021.3050055.

[16] 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, 2006.

[17] P. N. Belhumeur, J. P. Hespanha, and D. 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.

[18] 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.

[19] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.

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

[21] Z. Li, Z. Zhang, J. Qin, Z. Zhang, and L. Shao, "Discriminative Fisher embedding dictionary learning algorithm for object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 786–800, Mar. 2020.

[22] J. Wright, A. Y. Yang, A. Ganesh, S. S. 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.

[23] X. Wei *et al.*, "Reconstructible nonlinear dimensionality reduction via joint dictionary learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 175–189, Jan. 2019.

[24] 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.

[25] 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.

[26] 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.

[27] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[28] 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. (ICCV)*, Dec. 2013, pp. 689–696.

[29] 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.

[30] 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.

[31] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Discriminative multi-scale sparse coding for single-sample face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 302–312, Apr. 2017.

[32] 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.

[33] 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, May 2019, doi: 10.1109/TIFS.2019.2919950.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] H. Yan, J. Lu, X. Zhou, and Y. Shang, "Multi-feature multi-manifold learning for single-sample face recognition," *Neurocomputing*, vol. 143, pp. 134–143, Nov. 2014.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

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

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

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

[55] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

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

[57] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 5th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2002, pp. 53–58.

[58] 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.

[59] A. S. Georghiades, P. N. Belhumeur, and D. 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.

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

[61] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 947–954.
[62] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.
[63] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, Amherst, MA, USA, Tech. Rep. 07–49, Oct. 2007.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[65] J. Anderson, C. Otto, B. Maze, N. Kalka, and J. A. Duncan, "Understanding confounding factors in face detection and recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.
[66] D. Díaz-Vico and J. R. Dorronsoro, "Deep least squares Fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2752–2763, Aug. 2020.
[67] Q. Tan *et al.*, "Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 15, 2020, doi: 10.1109/TNNLS.2020.3025813.
[68] G. Zhang, M. Piccardi, and E. Z. Borzeshi, "Sequential labeling with structural SVM under nondecomposable losses," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4177–4188, Sep. 2018.
[69] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.
[70] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4432–4441.
[71] Y. Liu, Q. Sun, X. He, A.-A. Liu, Y. Su, and T.-S. Chua, "Generating face images with attributes for free," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2733–2743, Jun. 2021.
[72] 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.
[73] 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
[74] 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.
[75] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1891–1898.
[76] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, vol. 1, no. 3, p. 6.

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

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

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

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

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

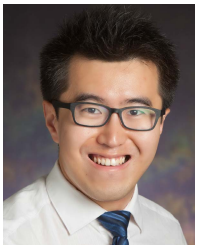
He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, pattern recognition, visual computing, and optimization.

Dr. Cheung is an Institution of Engineering and Technology (IET) Fellow, a British Computer Society (BCS) Fellow, a Royal Society of Arts (RSA) Fellow, and an International Engineering and Technology Institute (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, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2005.

After five years in the industry, he joined the University of Pittsburgh, Pittsburgh, PA, USA, in 2010, as an Assistant Professor, where he was promoted to an Associate Professor with tenure in 2014, and held a Bicentennial Alumni Faculty fellow position. He is currently a Professor with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, and serves as the Director of the NSF Industry-University Cooperative Research Center for Alternative Sustainable and Intelligent Computing and the Co-Director of the Duke Center for Computational Evolutionary Intelligence, focusing on the research of new memory and storage systems, machine learning and neuromorphic computing, and mobile computing systems.

Dr. Chen is a fellow of the ACM. 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 Systems Council (SYSC)/Council on Electronic Design Automation (CEDA) Technical Committee on Cyber-Physical Systems (TCCPS) Mid-Career Award. He is currently serving as the Editor-in-Chief for the *IEEE Circuits and Systems Magazine*. He is a Distinguished Lecturer of the 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, Champaign, IL, USA, in 2015 and 2018, respectively.

He was a Researcher with Dolby Laboratories, San Francisco, CA, USA. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computer vision, image and video processing, computational imaging, and big data applications.

Dr. Wen is currently a member of the IEEE Computational Imaging Technical Committee. He was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal of Singapore. His coauthored paper received the Top 10% Best Paper Award at the IEEE International Conference on Image Processing in 2014, and another paper received the Best Paper Runner-Up Award at the IEEE International Conference on Multimedia & Expo in 2020.