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

Meng Pang\textsuperscript{\textcopyright}, Binghui Wang\textsuperscript{\textcopyright}, Member, IEEE, Mang Ye\textsuperscript{\textcopyright}, Yiu-ming Cheung\textsuperscript{\textcopyright}, Fellow, IEEE, Yiran Chen\textsuperscript{\textcopyright}, Fellow, IEEE, and Bihan Wen\textsuperscript{\textcopyright}, 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,\textsuperscript{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\textsuperscript{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

\textsuperscript{1}More standardized biometric vocabularies can refer to the website of https://www.christoph-busch.de/standards.html

\textsuperscript{2}A prototype indicates a frontal face image with a neutral expression, under normal lighting, and without occlusion/disguise.
Re is featured in two aspects. Images in the pixel-spatial space, our proposed DisP for SSPP FR, as shown in Fig. 1(b). Compared with the classic enrolment database are contaminated by different facial variations. It is observed that the reconstructed image

linear representation classifier (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.

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.

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 = \( \{ D^{id}, D^{gan} \} \) and ˜D, where \( D^{id} \) and ˜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 ˜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 ˜D compete with each other such that G encodes a less discriminative variation feature, and meanwhile, generating the corresponding variation image that fools ˜D, i.e., G enables ˜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

3A 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 representation-based classification (S3RC) 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 S3RC 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 frontialization 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 sub-signals, 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 $y$, it can be represented as

$$ y = P\alpha + V\beta + e $$

where $P$, $V$, and $e$ are the enrolment sample dictionary, the variation dictionary, and small noise, respectively, $\alpha$ is the sparse coefficient vector, whose a few nonzero entries correspond to choosing a few numbers of enrolment samples (i.e., identities) from $P$, and $\beta$ is another sparse coefficient vector whose nonzero entries correspond to selecting a small subset of dictionary $V$. The coefficient vectors $\alpha$ and $\beta$ are calculated via solving the following optimization problem:

$$ \begin{align*}
\alpha^* = \arg \min_{\alpha} & \|y - [P \ V] \alpha\|^2_2 + \lambda \|\alpha\|_1 \\
\beta^* = \arg \min_{\beta} & \|y - [P \ V] \alpha + [V] \beta\|^2_2 + \lambda \|\beta\|_1 
\end{align*} $$

where $\lambda$ is a regularization parameter, $\| \cdot \|_1$ and $\| \cdot \|_2$ denote the $l_2$-norm and $l_1$-norm, respectively. Finally, similar to sparse representation-based classification (SRC) [22], $y$ 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 $x$ and the fake generated image $x^\prime$, while the generator $G$ is trained to generate realistic-looking images, i.e., $G(z)$, based on a random noise vector $z$ to fool $D$. Formally, the objective function of GAN is as follows:

$$ \max_G \min_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] $$

(1)

where $p_{data}$ and $p_z$ denote the distributions of the training data and the noise $z$, respectively. Alternatively, it has been shown that the minimization of $\mathbb{E}_{x \sim p_{data}} [\log (1 - D(G(z)))]$ can be replaced by the maximization of $\mathbb{E}_{z \sim p_z} [\log D(G(z))]$ to provide much stronger gradients early in learning [54]. Hence, the objective in (1) can be reformulated as follows:

$$ \max_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] $$

(2)

$$ \min_G \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log D(G(z))] $$

(3)

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 $x$, the proposed DisP+V aims to achieve the following objectives.

1) Disentangled Feature Learning: Learning a discriminative prototype feature $P(x)$ for $x$, such that $P(x)$: 1) represents the identity of $x$ and 2) is invariant to any facial variations in $x$; and learning a less discriminative variation feature $V(x)$ for $x$ such that $V(x)$ is irrelevant to the input identity information.

2) Prototype and Variation Generation: Recovering a high-quality (i.e., realistic looking) prototype $x^p$ for the input face image $x$, such that $x^p$: 1) is variation-free and 2) preserves the identity of $x$; and extracting the variation image $x^v$ (such that it: 1) captures the facial variation in $x$ and 2) contains little identity information of $x$.

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
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 \(G_{\text{enc}}\) and a decoder \(G_{\text{dec}}\). Given an input face image \(x\), \(G_{\text{enc}}\) has two separate branches, which aim to encode a more discriminative prototype feature \(P(x)\) and a less discriminative variation feature \(V(x)\) in a latent space. Subsequently, \(G_{\text{dec}}\) takes \(P(x)\) and \(V(x)\) as the inputs, and generates an appropriate prototype, i.e., \(x^p = G_{\text{dec}}(P(x))\), a variation image, i.e., \(x^v = G_{\text{dec}}(V(x))\), and a reconstructed image of \(x\), i.e., \(\tilde{x} = G_{\text{dec}}(f(x))\), respectively.

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

1) \(D_{\text{id}}\) outputs a \(N_{\text{id}}\)-dimensional vector for face identity classification, with \(N_{\text{id}}\) the total number of identities.

2) \(D_{\text{gan}}\) is a standard GAN discriminator to distinguish the real prototype versus fake prototype generated by the generator \(G\). More specifically, \(D_{\text{gan}}\) 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 \(N_{\text{id}}\)-dimensional vector and is used to predict the face identity label. Unlike \(D_{\text{id}}\) of \(D\), \(\tilde{D}\) only relates to the variation image \(x^v\).

2) DisP+V Training: Suppose we are given a training set of \(N_{\text{id}}\) identities with each face image \(x\) annotated by the label \(l = (l_{\text{id}}, l_{\text{var}})\), where \(l_{\text{id}}\) and \(l_{\text{var}}\) \((l_{\text{var}} = 1\ or \ 0)\) 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 \(l_{\text{var}}\) to form the real prototype corpus. We denote each standard/real prototype as \(x^p\), and its distribution as \(\mathcal{P}_{\text{real}}\), i.e., \(x^p \sim \mathcal{P}_{\text{real}}\). As a comparison, we denote that all face images \(x\) in the training set are sampled from the distribution \(\mathcal{P}_{\text{data}}\), i.e., \(x \sim \mathcal{P}_{\text{data}}\).

For the generator \(G\), we have the following four objectives.

1) \(D_{\text{id}}\) to classify the generated prototype \(x^p\) as the same identity label as the input image \(x\), i.e., \(l_{\text{id}}\).

2) \(D_{\text{gan}}\) to classify the generated fake prototype \(\tilde{x}^p\) as a real prototype, i.e., \(G\) enables \(D_{\text{gan}}\) to assign a high score to \(x^p\) of being real prototype.

3) \(\tilde{D}\) and make it fail to classify the generated variation \(x^v\), i.e., \(G\) enables \(\tilde{D}\) to output a constant vector with each element value equaling to \((1/N_{\text{id}})\).

4) \(\tilde{D}\) to well reconstruct the original input image \(x\).

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

\[
\max_G V_G = V_{\text{gan}}^G + \mu_1 V_{\text{id}}^G + \mu_2 V_{\text{var}}^G - \mu_3 V_{\text{rec}}^G \tag{6}
\]

where \(\mu_1, \mu_2,\) and \(\mu_3\) are three positive tradeoff parameters for the hybrid objective \(V_G\). The four subobjectives \(V_{\text{id}}^G, V_{\text{gan}}^G, V_{\text{var}}^G\), and \(V_{\text{rec}}^G\) are defined as follows:

\[
V_{\text{id}}^G (G, D, x) = E_x[\log D_{\text{id}}(G_{\text{dec}}(P(x)))] \tag{7}
\]

\[
V_{\text{gan}}^G (G, D, x) = E_x[\log D_{\text{gan}}(G_{\text{dec}} (P(x)))] \tag{8}
\]

\[
V_{\text{id}}^G (G, D, x) = H_x[\tilde{D}(G_{\text{dec}}(V(x)))] \tag{9}
\]

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

where \(D_{\text{id}}^i\) denotes the \(i\)th element in \(D_{\text{id}}\), \(H(\cdot)\) and \(||\cdot||_F\) denote the empirical entropy and the Frobenius norm, respectively. It is worth noting that maximizing entropy of the predicted identity distribution for \(G_{\text{dec}}(V(x))\), i.e., \(H_x[\tilde{D}(G_{\text{dec}}(V(x)))]\), in (9) is equivalent to the third objective that forces \(D\) 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 \(x\), \(D_{\text{id}}\) aims to correctly predict its identity label \(l_{\text{id}}\).

2) Given the real prototype \(x^p\) and the generated fake prototype by \(G\), i.e., \(x^v = G_{\text{dec}}(P(x))\), \(D_{\text{gan}}\) aims to classify \(x^p\) as the real prototype and classify \(x^v\) as the fake prototype.

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

\[
\max_D V_D = V_{\text{gan}}^D + \gamma V_{\text{id}}^D \tag{11}
\]

where \(\gamma\) is a positive tradeoff parameter, and \(V_{\text{id}}^D\) and \(V_{\text{gan}}^D\) are defined as follows:

\[
V_{\text{id}}^D (D, x) = E_x[\log D_{\text{id}}^D (x)] \tag{12}
\]

\[
V_{\text{gan}}^D (G, D_{\text{gan}}, x^p, x) = E_x[\log D_{\text{gan}}^D (x^p)] + E_x[\log (1 - D_{\text{gan}}^D (G_{\text{dec}}(P(x)))).] \tag{13}
\]

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

\[
\max_{\tilde{D}} V_{\tilde{D}} = E_x[\log \tilde{D}_{\text{id}}(G_{\text{dec}}(V(x)))] \tag{14}
\]

where \(\tilde{D}_{i}\) denotes the \(i\)th element in \(\tilde{D}\).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>Image in training set, (x \sim \mathcal{P}_{\text{data}})</td>
</tr>
<tr>
<td>(l_{\text{id}})</td>
<td>The identity label for (x)</td>
</tr>
<tr>
<td>(x^p)</td>
<td>Real prototype in training set, (x^p \sim \mathcal{P}_{\text{real}})</td>
</tr>
<tr>
<td>(G)</td>
<td>The encoder-decoder structural generator</td>
</tr>
<tr>
<td>(G_{\text{enc}})</td>
<td>The encoder in (G)</td>
</tr>
<tr>
<td>(P(x))</td>
<td>The learned prototype feature of (x)</td>
</tr>
<tr>
<td>(V(x))</td>
<td>The learned variation feature of (x)</td>
</tr>
<tr>
<td>(f(x))</td>
<td>The superposition of (P(x)) and (V(x))</td>
</tr>
<tr>
<td>(G_{\text{dec}})</td>
<td>The decoder in (G)</td>
</tr>
<tr>
<td>(x^p)</td>
<td>The generated prototype of (x)</td>
</tr>
<tr>
<td>(x^v)</td>
<td>The generated variation image of (x)</td>
</tr>
<tr>
<td>(\tilde{x})</td>
<td>The reconstructed image of (x)</td>
</tr>
<tr>
<td>(D)</td>
<td>The multi-task discriminator, i.e., (D = [D_{\text{gan}}, D_{\text{id}}])</td>
</tr>
<tr>
<td>(D_{\text{gan}})</td>
<td>The sub-discriminator to classify real and fake prototypes</td>
</tr>
<tr>
<td>(D_{\text{id}})</td>
<td>The sub-discriminator to predict identity label</td>
</tr>
<tr>
<td>(\tilde{D})</td>
<td>The identity discriminator only relates to (x^v)</td>
</tr>
<tr>
<td>(y)</td>
<td>Query image in testing set</td>
</tr>
<tr>
<td>(\mathcal{S})</td>
<td>The SSPP enrollment set in testing set</td>
</tr>
<tr>
<td>(P(y))</td>
<td>The learned prototype feature of (y)</td>
</tr>
<tr>
<td>(P(S))</td>
<td>The learned prototype features of (S)</td>
</tr>
</tbody>
</table>

**Table I**

**MEANING OF THE SYMBOLS IN DISP+V**
Algorithm 1 DisP+V Training

Input: A training set of $N_d$ identities with each image $x$ annotated by the label $l = [l^m, l^{var}]$; A real prototype corpus with each image $x^p$ sampled from the distribution $P_{real}$.

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

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

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 $\hat{D}$ by solving the objective functions $V_G$ in (6), $V_P$ in (11), and $V_{\hat{D}}$ in (14) iteratively. During the alternative training process, $G$, $D$, and $\hat{D}$ will be updated and improved. Specifically, with $D^{gem}$ in $D$ being more powerful in distinguishing real versus fake prototypes, $G$ strives for generating a realistic-looking prototype in order to fool $D^{gem}$. Besides, $\hat{D}^{id}$ in $\hat{D}$ 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 $\hat{D}$ being more powerful in classifying identity labels, $G$ makes efforts to capture the less discriminative characteristics (i.e., facial variations) in $x^o$ to fool $\hat{D}$ 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 $\hat{D}$ 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 $\hat{D}$ 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 $\hat{D}$, while not directly using $D^{id}$, to predict the identity label for the generated variation $x^o$. This strategy reduces the training complexity for $D^{id}$ and enables $D^{id}$ and $\hat{D}$ to be responsible for their respective adversarial learning.

C. Application Scenarios

1) SSPP FR: Let $y$ be a new query sample from the testing set, $S = [s_1, \ldots, s_n]$ be the SSPP enrolment set with $n$ identities, and $A = [a_1, \ldots, a_m]$ be the generic set with $m$ samples from other identities not of interest. With the trained DisP+V model, we can obtain the prototype feature of $y$, i.e., $P(y)$, and the prototype features of $S$, i.e., $P(S)$. Subsequently, we classify the identity of $y$ by matching $P(y)$ with $P(S) = [P(s_1), \ldots, P(s_n)]$ as follows:

Scheme 1: $ID(y) = \arg\min_k \text{dist}(P(y), P(s_k))$ (15)

where $\text{dist}(a, b)$ represents the distance between the feature vectors of $a$ and $b$, and the arc-cosine-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 $y$, i.e., $f(y) = P(y) + V(y)$, and the variation features of $A$, i.e., $V(A)$. Then, we solve the following $l_1$-based optimization problem:

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

(16)

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

Scheme 2: $ID(y) = \arg\min_k r_k(y)$ (17)

where $r_k(y)$ is computed by

$$r_k(y) = \left\| f(y) - [P(S) V(A)] [\delta_k(\alpha^*)] \right\|_2^2$$

(18)

where $\delta_k(\alpha^*)$ being a vector whose nonzero entries are the entries in $\alpha^*$ 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+Vpv hereinafter. Furthermore, we analyze the time complexities of DisP+V and DisP+Vpv for recognizing the query sample $y$, respectively. Specifically, both of the recognition stages for DisP+V and DisP+Vpv include two steps: feature extraction and classification. Suppose the image size of $y$ is $w \times h$ and the number of the convolutional layers in $G_{enc}$ is $L$. The time complexities of DisP+V and DisP+Vpv 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 $P(y)$ and $n$ is the size of the enrolment set $S$; and the time complexity of DisP+Vpv is $O(\tau d^2 + \tau d(n + q))$ [8] where $\tau$ is the number of iterations for BPDN-homotopy in. (16) and $q$ is the size of the generic set $A$. Overall, the time complexities of DisP+V and DisP+Vpv 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+Vpv 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+Vv 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 $D^1$, $D^2$, and $D$ 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 100 identities are randomly chosen for training and the rest 108 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 99594 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 Alex 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 50000 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 130000 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>
<thead>
<tr>
<th>Layer</th>
<th>$G_{enc}$</th>
<th>$G_{dec}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>3 x 3 / 1 / 1</td>
<td>96 x 96 x 32</td>
</tr>
<tr>
<td>Conv2</td>
<td>3 x 3 / 1 / 1</td>
<td>96 x 96 x 64</td>
</tr>
<tr>
<td>Conv3</td>
<td>3 x 3 / 2 / 0</td>
<td>48 x 48 x 64</td>
</tr>
<tr>
<td>Conv4</td>
<td>3 x 3 / 1 / 1</td>
<td>48 x 48 x 64</td>
</tr>
<tr>
<td>Conv5</td>
<td>3 x 3 / 2 / 0</td>
<td>48 x 48 x 128</td>
</tr>
<tr>
<td>Conv6</td>
<td>3 x 3 / 1 / 1</td>
<td>24 x 24 x 128</td>
</tr>
<tr>
<td>Conv7</td>
<td>3 x 3 / 1 / 1</td>
<td>24 x 24 x 96</td>
</tr>
<tr>
<td>Conv8</td>
<td>3 x 3 / 1 / 1</td>
<td>24 x 24 x 192</td>
</tr>
<tr>
<td>Conv9</td>
<td>3 x 3 / 2 / 0</td>
<td>12 x 12 x 192</td>
</tr>
<tr>
<td>Conv10</td>
<td>3 x 3 / 1 / 1</td>
<td>12 x 12 x 256</td>
</tr>
<tr>
<td>Conv11</td>
<td>3 x 3 / 1 / 1</td>
<td>12 x 12 x 256</td>
</tr>
<tr>
<td>Conv12</td>
<td>3 x 3 / 2 / 0</td>
<td>6 x 6 x 256</td>
</tr>
<tr>
<td>Conv13</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 160</td>
</tr>
<tr>
<td>Conv14</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 320</td>
</tr>
<tr>
<td>Conv15-1, Conv15-2</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 320</td>
</tr>
<tr>
<td>Conv16-1, Conv16-2</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 320</td>
</tr>
<tr>
<td>Conv17-1, Conv17-2</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 256</td>
</tr>
<tr>
<td>Pool18-1, Pool18-2</td>
<td>6 x 6 / 1 / 0</td>
<td>1 x 1 x 256</td>
</tr>
</tbody>
</table>

Authorized licensed use limited to: Hong Kong Baptist University. Downloaded on February 05,2023 at 08:06:42 UTC from IEEE Xplore. Restrictions apply.
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

<table>
<thead>
<tr>
<th>Layer</th>
<th>Filter / Stride / Pad</th>
<th>Output Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>3 x 3 / 1 / 1</td>
<td>96 x 96 x 32</td>
</tr>
<tr>
<td>Conv2</td>
<td>3 x 3 / 1 / 1</td>
<td>96 x 96 x 64</td>
</tr>
<tr>
<td>Conv3</td>
<td>3 x 3 / 2 / 0</td>
<td>48 x 48 x 64</td>
</tr>
<tr>
<td>Conv4</td>
<td>3 x 3 / 1 / 1</td>
<td>48 x 48 x 64</td>
</tr>
<tr>
<td>Conv5</td>
<td>3 x 3 / 1 / 1</td>
<td>48 x 48 x 128</td>
</tr>
<tr>
<td>Conv6</td>
<td>3 x 3 / 2 / 0</td>
<td>24 x 24 x 128</td>
</tr>
<tr>
<td>Conv7</td>
<td>3 x 3 / 1 / 1</td>
<td>24 x 24 x 96</td>
</tr>
<tr>
<td>Conv8</td>
<td>3 x 3 / 1 / 1</td>
<td>24 x 24 x 192</td>
</tr>
<tr>
<td>Conv9</td>
<td>3 x 3 / 2 / 0</td>
<td>12 x 12 x 192</td>
</tr>
<tr>
<td>Conv10</td>
<td>3 x 3 / 1 / 1</td>
<td>12 x 12 x 128</td>
</tr>
<tr>
<td>Conv11</td>
<td>3 x 3 / 1 / 1</td>
<td>12 x 12 x 256</td>
</tr>
<tr>
<td>Conv12</td>
<td>3 x 3 / 2 / 0</td>
<td>6 x 6 x 256</td>
</tr>
<tr>
<td>Conv13</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 160</td>
</tr>
<tr>
<td>Conv14</td>
<td>3 x 3 / 1 / 1</td>
<td>6 x 6 x 320</td>
</tr>
<tr>
<td>Pool15</td>
<td>6 x 6 / 1 / 0</td>
<td>1 x 1 x 320</td>
</tr>
<tr>
<td>FC16(D)</td>
<td>-</td>
<td>( N_d+1 )</td>
</tr>
<tr>
<td>FC16(D̃)</td>
<td>-</td>
<td>( N_\tilde{d} )</td>
</tr>
</tbody>
</table>

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 \( G_{\text{enc}} \) and \( G_{\text{dec}} \), where batch normalization and exponential linear unit are used after each \( \text{conv} \) and \( \text{deconv} \) layer. In \( G_{\text{enc}} \), the final \text{AvgPool} layer is replaced by two subnets with each having three \text{conv} layers and one global \text{AvgPool}. The two disentanglement branches extract two 256-D features for \( x \), i.e., \( P(x) \) and \( V(x) \). Subsequently, \( P(x) \), \( V(x) \), and their superposition, i.e., \( f(x) = P(x) + V(x) \), are used as the inputs for \( G_{\text{dec}} \) to generate the prototype \( \tilde{x} \), the variation image \( x' \), and the reconstructed image \( \tilde{x} \) for \( x \), 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 \( D_{\text{id}} \) for predicting the face identity and the rest one is reserved for \( D_{\text{sim}} \) to distinguish real versus fake prototype. The output of \( \tilde{D} \) is a \( N_{\tilde{d}} \)-dimensional vector only for face identity prediction. The network structures of \( D \) and \( \tilde{D} \) are presented in Table III.

We train the proposed DisP+V4 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 achieves promising performance when the tradeoff parameters \( \mu_1, \mu_2, \) and \( \mu_3 \) in (6) and \( \gamma \) 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.
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 S3RC [43] method. For SVDL, SLRC, and S3RC, the training set is used as the auxiliary generic set for generating variation dictionaries. We tune the regularization parameter $\lambda$ of SRC, SLRC, and S3RC and find that they achieve the best performance when $\lambda = 0.01$. For SVDL, as suggested in [28], the parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ are set at 0.001, 0.01, and 0.0001, respectively. For DisP+Vpv, the regularization parameter $\lambda$ 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 ($\pm$ standard errors) and the recognition time (s) 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+Vpv 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+Vpv 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+Vpv 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+Vpv 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.

It is interesting to find that each of DisP+V (based on direct prototype feature matching) and DisP+Vpv (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+Vpv is better at handling additive variations, such as disguise on CAS-PEAL.

S3RC 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, S3RC 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 S3RC depends heavily on the clustering performance of GMM, which is sensitive to severe lightings and shadows.

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.

Although the representation learning-based VAE also performs variation disentanglement during encoding, it is much less competitive with our DisP+V and DisP+Vpv. 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+Vpv, which is consistent with the complexity analysis results in Section IV-C. Moreover, the recognition time of DisP+V and DisP+Vpv 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. S3RC 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.

### 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TPR(%)@FAR=0.1</th>
<th>AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>DisP+V</td>
<td>Baseline</td>
</tr>
<tr>
<td>Multi-PIE [Expression]</td>
<td>60.8±3.1</td>
<td>61.0±3.1</td>
</tr>
<tr>
<td>FERET [Pose]</td>
<td>44.6±3.1</td>
<td>52.4±2.2</td>
</tr>
<tr>
<td>CAS-PEAL [Disguise]</td>
<td>49.7±3.2</td>
<td>56.7±3.2</td>
</tr>
<tr>
<td>E-Yale-B&amp;AR Light [Lighting]</td>
<td>63.9±2.2</td>
<td>80.0±0.8</td>
</tr>
<tr>
<td>FRGC v2.0 [Multiple variations]</td>
<td>78.0±2.5</td>
<td>83.2±1.7</td>
</tr>
</tbody>
</table>

**Table VI**

Verification Performance of DisP+V on the Multi-PIE, FERET, CAS-PEAL, E-Yale-B&AR Light, and FRGC v2.0 Datasets
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 (± 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., $D = [D^{id}, D^{gan}]$ 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.

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 SSP-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 $x^{v}$ individually. In the experiment, we remove $\tilde{D}$ and directly use $D^{id}$ to predict both IDs of $x^{p}$ and $x^{v}$ and perform two adversarial learning based on $D^{id}$. We denote the DisP+V variant as DisP+V w/o $D^{id}$ 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. 5.** Comparison results of DisP+V and its variants DisP+V w/o $D^{id}$ and DisP+V w/o $D^{gan}$ 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 $D^{gan}$, and the true prototype for reference, respectively.

**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.
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

<table>
<thead>
<tr>
<th>Method</th>
<th>SSPP-se FR</th>
<th>SSPP-ce FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRC</td>
<td>50.2</td>
<td>47.3±1.5</td>
</tr>
<tr>
<td>SVDL</td>
<td>57.6</td>
<td>52.9±1.6</td>
</tr>
<tr>
<td>SLRC</td>
<td>65.1</td>
<td>62.6±2.5</td>
</tr>
<tr>
<td>S3RC</td>
<td>65.6</td>
<td>62.9±2.3</td>
</tr>
<tr>
<td>DisP+V</td>
<td>72.2</td>
<td>70.9±1.6</td>
</tr>
</tbody>
</table>

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 S3RC. 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 $G_{\text{dec}}$ 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, $G_{\text{dec}}$, $D$, and $\tilde{D}$. We denote our DisP+V using the LightCNN-29 feature extractor as DisP+VLC29, 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], regular-face [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+VLC29 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 S3RC in Table VII and that of ordinary glasses, sunglasses, and hats successively, and the corresponding interpolated images look natural.
TABLE VIII

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepID</td>
<td>70.7</td>
</tr>
<tr>
<td>JCR-ACF</td>
<td>86.0</td>
</tr>
<tr>
<td>VGG-face</td>
<td>84.7</td>
</tr>
<tr>
<td>Regular-face</td>
<td>83.7</td>
</tr>
<tr>
<td>Arc-face</td>
<td>92.3</td>
</tr>
<tr>
<td>CJR-RACF</td>
<td>95.5</td>
</tr>
<tr>
<td>DisP+V_{LC29}</td>
<td>96.7</td>
</tr>
</tbody>
</table>

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


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 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 multimedia retrieval, 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.

Authorized licensed use limited to: Hong Kong Baptist University. Downloaded on February 05,2023 at 08:06:42 UTC from IEEE Xplore. Restrictions apply.