IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 16, 2020, accepted July 25, 2020, date of publication August 7, 2020, date of current version August 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014877

Dual-View Normalization for Face Recognition

GEE-SERN HSU, (Senior Member, IEEE), AND CHIA-HAO TANG, (Student Member, IEEE)
Artificial Vision Laboratory, Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan

Corresponding author: Gee-Sern Hsu (jison@mail.ntust.edu.tw)

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan, under Grant 106-2221-E-011-144-MY3, and in part by the Center for Cyber-Physical System Innovation through the Higher Education Sprout Project, the Ministry of Education (MOE), Taiwan.

ABSTRACT Face normalization refers to a family of approaches that rotate a non-frontal face to the frontal pose for better handling of face recognition. While a great majority of face normalization methods focus on frontal pose only, we proposed a framework for dual-view normalization that generates a frontal pose and an additional yaw-45° pose to an input face of an arbitrary pose. The proposed Dual-View Normalization (DVN) framework is designed to learn the transformation from a source set to two normal sets. The source set contains faces collected *in the wild* and covers a wide scope of variables. One normal set contains face images taken under controlled conditions and all faces are in frontal pose and balanced in illumination. The other normal set contains faces also taken under controlled conditions and balanced in illumination, but in 45° pose. The DVN framework is composed of one face encoder, two layers of generators, and two sets of discriminators. The encoder is made of a state-of-the-art face recognition network, which is not updated during training, and it acts as a facial feature extractor. The Layer-1 generators are trained on both the source and normal sets, aiming at learning the transformation from the source set to both normal sets. The trained generators can transform an arbitrary face into a pair of normalized faces, one in frontal pose and the other in 45° pose. The Layer-2 generators are trained to enhance the identity preservation of the faces made by the Layer-1 generators by minimizing the cross-pose identity loss. The discriminators are trained to ensure the photo-realistic quality of the dual-view normalized face images generated by the generators. The loss functions employed in the generators and the discriminators are designed to achieve satisfactory dual-view normalization outcomes and identity preservation. We verify the DVN framework on benchmark databases and compare with other state-of-the-art approaches for tackling face recognition.

INDEX TERMS Face recognition, face normalization, face synthesis.

I. INTRODUCTION

Face normalization aims to normalize a non-frontal face by rotating it back to the frontal pose for achieving better recognition performance. While a great majority of face normalization methods focus on frontal pose only, we propose a framework for dual-view normalization that generates a frontal pose and an additional yaw-45° pose to a face of an arbitrary pose. The proposed Dual-View Normalization (DVN) framework is designed to learn the transformation from a source set to two normal sets. The source set contains faces with a wide range of variation across illumination, pose, expression and other variables. One normal set contains face images taken under controlled conditions and all faces are in frontal pose and balanced in illumination. The other normal set contains faces also taken under controlled conditions and balanced in illumination, but in 45° pose. The application scope of the proposed approach is not limited to face recognition, it can be applied for face synthesis, e.g., making a required ID photo from a daily-life photo.

The proposed DVN, shown in Figure. 1, is composed of a face encoder, two pairs of generators, and two sets of discriminators. The face encoder is built on the feature embedding layers of a state-of-the-art face recognition network, the ArcFace [1], and serves as a facial feature extractor that can offer consistent features regardless of illumination, pose, expression and other variables. The Layer-1 generators are trained on the source set and normal sets, aiming at learning the transformation from the source domain to the normal domain and preserving facial identities. After the learning on the dual-view data, it can transform an arbitrary face into a pair normalized faces, one in frontal-view pose and the other

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

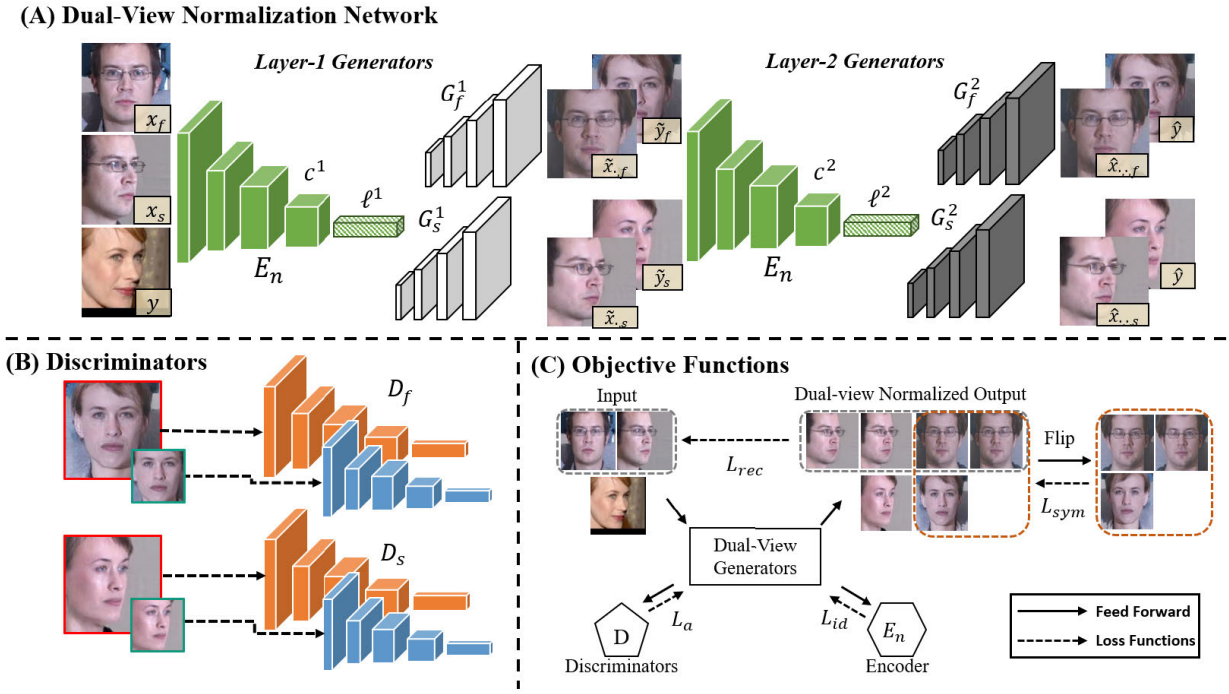

FIGURE 1. The configuration of the proposed Dual-View Normalization (DVN) framework. It is composed of one face encoder E_n, two layers of generators $[G_f^1, G_s^1]$ and $[G_f^2, G_s^2]$ and two sets of discriminators $[D_{f,k}^1, D_{s,k}^1]_{k=1,2}$ and $[D_{f,k}^2, D_{s,k}^2]_{k=1,2}$. It is trained on the source set, composed of faces collected in the wild, and two normal sets, composed of frontal- and 45° side-view faces.

in 45° side-view pose. The Layer-2 generators are trained on the source set, the normal sets, and the generated data from Layer-1 generators, aiming to enhance the identity preservation of the faces made by Layer-1 generators by minimizing the cross-pose identity loss. The discriminators are trained to ensure the normalized views and the photo-realistic quality of the generated face images. The loss functions employed in the generators and the discriminators are appropriately designed for achieving better normalization outcomes and recognition performance.

Most face normalization approaches consider the frontal pose as the only target when transforming a non-frontal face. The Two-Pathway Generative Adversarial Network (TP-GAN) [2] adopts a two-pathway architecture along with various loss functions to synthesize the frontal view of an arbitrary input face. The Pose Invariant Model (PIM) [3] proposes an end-to-end unified architecture and a ''learning to learn'' strategy for identity-preserving frontal-view synthesis. The High Fidelity Pose Invariant Model (HF-PIM) [4] imposes texture warping and leverages a dense correspondence field to bind the 2D and 3D surface spaces. Experiments show that the HF-PIM can improve high-resolution frontalized appearances. The Face Normalization Model (FNM) [5] employs a face expert network to be part of the generator for retaining face identity. With five local-region based discriminators, the FNM can transform an arbitrary face to the frontal pose with identity preserved. The face encoder in the proposed DVN is similar to the face expert in the FNM, which explores a pretrained face recognition network for face encoding. As face encoding is an important part of the framework, both the DVN and FNM do not build an additional face encoding module, and instead employ the latest face recognition networks. This step substantially reduces the complexity of the framework, making the computation highly efficient.

The primary novelty of the proposed DVN is the dual-view normalization, which considers the yaw-45° side-view pose as another normal view in addition to the common frontal-view. Similar to the forensic facial records, both frontal and profile face images are kept for better visual perception of the subject. The additional ''normal'' 45° side view is verified, through our experiments, to be able to better represent a given face and yield better cross-pose recognition. Two reasons are given below for selecting the 45° side-view pose as the additional normal view instead of the 90° profile pose:

- The face encoder in the DVN framework knows better of the faces within 45° as the majority of the training data are within this pose range;
- The facial region of interest, i.e., the region with most facial traits, reduces to some minimal level at profile pose which will degrade the recognition performance. This is verified in our experiments. As the DVN allows the specification of the normal view, we compare the performance of using the side view as the normal view and that by using the profile view. The former leads to a better performance on a benchmark database.

The contributions of this work can be summarized as:

- Different from the existing face normalization approaches which consider the frontal pose as the only normalized view, the proposed DVN framework defines normalization in dual views, one in frontal view and the other in yaw-45° side view. The additional side-view normalization can better represent a face, leading to a better recognition performance.
- The proposed DVN integrates identity preservation, face normalization and pose transformation so that it can transform a face of an arbitrary pose to specific normalized poses with identity well retained.
- The proposed DVN is verified highly competitive to state-of-the-art methods for face recognition.

The code is available at https://github.com/HaoRecog/Dual-view-Normalization-for-Face-Recognition. In the following, we first give a review to the state-of-the-art approaches for face normalization in Sec. II. The proposed Dual-View Normalization (DVN) framework is presented in Sec. III, followed by the experiments for qualitative and quantitative evaluations reported in Sec. IV. A conclusion to this work is given in Sec. V.

II. RELATED WORK

Many recent studies [2]–[5] demonstrate the advantages of using face normalization to tackle face recognition, as a face of an arbitrary pose can be normalized to the frontal pose for better feature extraction. The TP-GAN, proposed by Huang *et al.* [2], adopts a two-pathway architecture to achieve frontal-view synthesis. The adversarial loss, symmetry loss and identity preserving loss are combined to guide an identity-preserving inference of the frontal views from the profile views. The Face Frontalization Generative Adversarial Network (FF-GAN), proposed by Yin *et al.* [6], incorporates the 3DMM [7] into the GAN framework, which exploits the prior knowledge in shape and appearance. The 3DMM-conditioned GAN can generate an image that maintains both the global pose accuracy and local characteristics for improved frontalization. Based on the TP-GAN, Zhao *et al.* [3] proposed the Pose Invariant Model (PIM) to perform face normalization and extract pose invariant features jointly, and make the two tasks benefit to each other. The PIM incorporates an unsupervised cross-domain adversarial training and a ''learning to learn'' strategy for identity preservation.

The HF-PIM, proposed by Cao *et al.* [4], introduces the dense correspondence field to bind the 2D and 3D surface spaces. The method decomposes the warping procedure into dense correspondence field estimation and facial texture map recovering, improving the high-resolution frontalization and pose-invariant face recognition. Qian *et al.* [5] proposed the Face Normalization Model (FNM), which employs a face expert network in the generator to monitor identity preservation while exploiting the pixel-wise loss to stabilize optimization process. To refine the texture, a series of local attention discriminators with fixed areas are explored for synthesizing regional details. The CAPG-GAN [8] extracts the head pose information from the facial landmark heatmaps for not only forming a mask to guide the generator in making images, but also providing a flexible controllable condition for improving the image quality. The Disentangled Representation-learning GAN (DR-GAN) [9], [10] considers face normalization as disentangled representation learning, and proposed a framework to learn the disentangled facial representation on top of the face image generation. An *et al.* [11] proposed the Adaptive Pose Alignment (APA) method to learn multiple pose-specific templates for face alignment, and a feature normalization technique to generate discriminative facial representation combined with the APA.

To preserve the facial identity across normalization, many of the above referred works consider the identity latent codes extracted from face images. Most of them train an encoder to transform an input face into an identity latent code, which is usually processed by minimizing the identity loss and then decoded into a frontalized face [2]–[4], [9], [10]. However, the proposed DVN explores the encoder made of the feature embedding layers of the ArcFace network [1], which does not require any training and provides the reliable pose-invariant facial representation needed for identity preservation.

As the proposed approach deals with the normalization of faces of arbitrary poses, the head pose estimation may also be considered as part of related work. A recent approach proposed by Barra *et al.* [12] first detects the facial landmarks, and applies the web-shaped model to associate each landmark to a specific face sector. The obtained information is used to build a feature vector to infer the head pose. Yuan *et al.* [13] propose a geometry-based method to estimate the head pose from a 2D face image. The head pose is obtained by minimizing the Euclidean distance between the normalized 2D feature points obtained from the non-coplanar feature points on a predefined 3D face model and the 2D re-projections of the morphed 3D feature points from the spherical coordinates. A multitask CNN is proposed by Elharrouss *et al.* [14] for handling the pose estimation and face recognition in a unified framework. Due to the pose classification enabled by the pose estimation, the faces in a particular pose class can be better recognized in a pose-oriented way. However, the above reviewed methods and most pose estimation approaches aim at precisely estimating the 3D pose of a 2D face image instead of normalizing non-frontal poses. The latter is the focus of this article.

III. PROPOSED APPROACH

The configuration of the proposed Dual-View Normalization (DVN) framework is illustrated in Figure 1. It is composed of a face encoder E_n, two layers of generators $[G_f^1, G_s^1]$ $[G_f^2, G_s^2]$, and two sets of discriminators $[D_{f,k}^1, D_{s,k}^1]_{k=1,2}$ and $[D_{f,k}^2, D_{s,k}^2]_{k=1,2}$, and is trained on a source set and two normal sets. The source set contains faces collected in the wild, and covers a wide range of variation across illumination, pose, expression, resolution, occlusion and other variables.

One normal set contains face images taken under controlled conditions, and all faces are in frontal pose and balanced in illumination. The other normal set contains faces also taken under controlled conditions and balanced in illumination, but in 45° pose. The face encoder E_n is made of the feature embedding layers of the ArcFace [1], which is a state-of-the-art CNN for face recognition. E_n is not updated during training and kept as the original from the trained ArcFace. The encoder E_n can transform an input image x to a feature representation c ($7 \times 7 \times 512$-dims) and a latent vector ℓ (1×512-dims), as shown in Figure 1. The feature representation c will be decoded to a pair of dual-view normalized faces $(\tilde{x}_f, \tilde{x}_s)$, where $\tilde{x}_f$ is the frontal-view normalized face and $\tilde{x}_s$ is the 45° side-view normalized face, by the Layer-1 generators G_f^1 and G_s^1, respectively (f stands for frontal view and s for side view). The latent vector ℓ will be exploited to compute the identity loss and guarantee the identity preservation (e.g., x_f and its corresponding synthetic faces $\tilde{x}_{f,f}$ and $\tilde{x}_{f,s}$). G_f^1 and G_s^1 are trained to make the dual-view face pair $(\tilde{x}_f, \tilde{x}_s)$ maintain the same identity as of x and appear in the required frontal and side-view poses. The Layer-2 generators G_f^2 and G_s^2 are trained to rotate the dual-view face pair $(\tilde{x}_f, \tilde{x}_s)$ to another dual-view face pair $(\hat{x}_f, \hat{x}_s)$ for enhancing the identity preservation across the pose transformation. The two sets of discriminators $[D_{f,k}^1, D_{Gs,k}^1]_{k=1,2}$ and $[D_{f,k}^2, D_{s,k}^2]_{k=1,2}$ are trained to force the generated images at the Layer-1 and Layer-2, respectively, to appear in the specified dual poses with photo-realistic details. In the above notations, the superscript ^{1} (^{2}) denotes the components at Layer-1 (2).

A. IDENTITY PRESERVING GENERATORS

The Layer-1 generators G_f^1 and G_s^1 are trained to generate the dual-view normalized faces $\tilde{x}_f$ and $\tilde{x}_s$ that keep the same identity as of the input x and appear in the defined specific poses, i.e., $\tilde{x}_f$ in frontal pose and $\tilde{x}_s$ in 45° pose. The training set has three subsets, the source set S_m and the pair of normal sets $(S_{n,f}, S_{n,s})$. S_m contains faces collected in the wild and exhibit a wide scope of variables. $S_{n,f}$ consists of frontal-view normal faces and $S_{n,s}$ consists of a 45° side-view normal faces. The normal faces refer to the face images collected in a controlled condition with balanced illumination and natural facial expression. For the conciseness of notation, the 45° side-view face will be referred as "side-view face" in the rest of the paper.

During training, given a face x_f from the frontal-view normal set $S_{n,f}$, the generator G_f^1 is trained to generate a frontal normalized face, i.e. $\tilde{x}_{f,f} = G_f^1(E_n(x_f))$, and G_s^1 is trained to generate a side-view normalized face, i.e., $\tilde{x}_{f,s} = G_s^1(E_n(x_f))$. The identity loss considers the minimization of the 2-norm between the latent vectors of the input x_f and of the generated $\tilde{x}_{f,f}, \tilde{x}_{f,s}$, and the latent vectors are obtained by passing the image x_f and $\tilde{x}_{f,f}, \tilde{x}_{f,s}$ thru the encoder E_n. Thus the identity loss is formulated as: (other losses are also considered in the training, and will be discussed in Sec. III-C. In this section, we focus on the identity loss first)

$$||E_n(x_f) - E_n(\tilde{x}_{f,f})||_2 + ||E_n(x_f) - E_n(\tilde{x}_{f,s})||_2 \quad (1)$$

Similarly considering a side-view face x_s from the side-view normal set $S_{n,s}$, the training considers the minimization of the following loss:

$$||E_n(x_s) - E_n(\tilde{x}_{s,f})||_2 + ||E_n(x_s) - E_n(\tilde{x}_{s,s})||_2 \quad (2)$$

When considering a face y from the source set S_m, the training also performs a similar loss as follows:

$$||E_n(y) - E_n(\tilde{y}_f)||_2 + ||E_n(y) - E_n(\tilde{y}_s)||_2 \quad (3)$$

In summary, when considering the identity loss, the training of G_f^1 and G_s^1 is performed by the same form of minimization, as shown in (1), (2) and (3), no matter where the training data come from.

Similar identity loss can be derived for the Layer-2 generators, but with more pose transformations included. The training set consists of 1) all the synthetic faces $[\tilde{x}]$, $[\tilde{y}]$ made by G_f^1 and G_s^1, 2) the normal sets $S_{n,f}$ and $S_{n,s}$ and 3) the source set S_m.

The Layer-2 generators G_f^2, G_s^2 are trained to fulfill two purposes. One is the enforcement of the identity preservation by taking the synthetic faces, generated by the G_f^1, G_s^1, as input and make the generated output faces identity-preserving. The other is the transformation between the two normalized poses as the frontal-view faces $[\tilde{x}_{f,f}, \tilde{x}_{s,f}, \tilde{y}_f]$ will be transformed to the side-view faces by G_s^2 and the side-view faces $[\tilde{x}_{s,s}, \tilde{x}_{f,s}, \tilde{y}_s]$ will be transformed to the frontal-view faces by G_f^2. The main difference from G_f^1, G_s^1 is that the G_f^2, G_s^2 need to consider the synthetic faces generated by G_f^1, G_s^1 as part of input, and generate the output faces to preserve the identity as of the original input to G_f^1, G_s^1.

When the training data $x_f \in S_{n,f}$ enters G_f^1, G_s^1, and then G_f^2, G_s^2, the generated data are $\hat{x}_{f,f,f}$, $\hat{x}_{f,f,s}$, $\hat{x}_{f,s,f}$ and $\hat{x}_{f,s,s}$, and the ID-loss can be written as follows:

$$\begin{aligned} &||E_n(x_f) - E_n(\hat{x}_{f,f,f})||_2 + ||E_n(x_f) - E_n(\hat{x}_{f,f,s})||_2 \\ &\quad + ||E_n(x_f) - E_n(\hat{x}_{f,s,f})||_2 + ||E_n(x_f) - E_n(\hat{x}_{f,s,s})||_2 \end{aligned} \quad (4)$$

where $\hat{x}_{f,f,f}$ and $\hat{x}_{f,s,f}$ denote the frontal-view normalized faces made by G_f^2 with $\tilde{x}_{f,f}$ and $\tilde{x}_{f,s}$ as the input respectively; $\hat{x}_{f,f,s}$ and $\hat{x}_{f,s,s}$ denote the side-view normalized faces made by G_s^2 with same input as above. Similarly, considering the training data $x_s \in S_{n,s}$ entering the Layer-1 and Layer-2 generators, the identity loss can be written as:

$$\begin{aligned} &||E_n(x_s) - E_n(\hat{x}_{s,f,f})||_2 + ||E_n(x_s) - E_n(\hat{x}_{s,s,f})||_2 \\ &\quad + ||E_n(x_s) - E_n(\hat{x}_{s,f,s})||_2 + ||E_n(x_s) - E_n(\hat{x}_{s,s,s})||_2 \end{aligned} \quad (5)$$

When considering the training data y entering the Layer-1 and Layer-2 generators, the identity loss can be written as:

$$\begin{aligned} &||E_n(y) - E_n(\hat{y}_{f,f})||_2 + ||E_n(y) - E_n(\hat{y}_{f,s})||_2 \\ &\quad + ||E_n(y) - E_n(\hat{y}_{s,f})||_2 + ||E_n(y) - E_n(\hat{y}_{s,s})||_2 \end{aligned} \quad (6)$$

Note that the losses in (4), (5), and (6) are for the training data entering the network from Layer 1. When the training data in $S_{n,f}$, $S_{n,s}$ and S_m directly enters the G_f^2, G_s^2, the identity losses considered are in the same forms as those in (1), (2), and (3) but modified for G_f^2, G_s^2.

In summary, if we denote the overall identity loss for training the Layer-1 generators G_f^1, G_s^1 as L_{id}^1 and that for training Layer-2 generators G_f^2, G_s^2 as L_{id}^2, L_{id}^1 is the summation of (1), (2), (3) and L_{id}^2 is the summation of (4), (5), (6), but (1), (2), and (3) modified for G_f^2, G_s^2.

TABLE 1. The network settings of the generators and discriminators in the DVN. Same settings for all generators, $D_{k=1}$ and $D_{k=2}$ refer to the global region and local region discriminators. The images that enter $D_{k=1}$ are sized 224 × 224 × 3, and the local-region ones that enter $D_{k=2}$ are sized 150 × 156 × 3.

G			$D_{k=1}$		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
Conv1	$1 \times 1/1$	$7 \times 7 \times 512$	Conv1	$3 \times 3/2$	$112 \times 112 \times 32$
ResBlock11	$3 \times 3/1$	$7 \times 7 \times 512$	Conv2	$3 \times 3/2$	$56 \times 56 \times 64$
ResBlock12	$3 \times 3/1$	$7 \times 7 \times 512$	Conv3	$3 \times 3/2$	$28 \times 28 \times 128$
ResBlock13	$3 \times 3/1$	$7 \times 7 \times 512$	Conv4	$3 \times 3/2$	$14 \times 14 \times 256$
ResBlock14	$3 \times 3/1$	$7 \times 7 \times 512$	Conv5	$3 \times 3/2$	$7 \times 7 \times 256$
FConv2	$4 \times 4/2$	$14 \times 14 \times 256$	FC		1
ResBlock2	$3 \times 3/1$	$14 \times 14 \times 256$	$D_{k=2}$		
FConv3	$4 \times 4/2$	$28 \times 28 \times 128$	Conv2	$3 \times 3/2$	$75 \times 78 \times 32$
ResBlock3	$3 \times 3/1$	$28 \times 28 \times 128$	Conv3	$3 \times 3/2$	$38 \times 39 \times 64$
FConv4	$4 \times 4/2$	$56 \times 56 \times 64$	Conv4	$3 \times 3/2$	$19 \times 20 \times 128$
ResBlock4	$3 \times 3/1$	$56 \times 56 \times 64$	Conv4	$3 \times 3/2$	$10 \times 10 \times 256$
FConv5	$4 \times 4/2$	$112 \times 112 \times 32$	FC		1
ResBlock5	$3 \times 3/1$	$112 \times 112 \times 32$			
FConv6	$4 \times 4/2$	$224 \times 224 \times 32$			
ResBlock6	$3 \times 3/1$	$224 \times 224 \times 32$			
Conv7	$1 \times 1/1$	$224 \times 224 \times 3$			
tanh					

The network settings of the generators are given in the left part of Table 1. The Layer-1 and the Layer-2 generators have the same architecture, however, they are completely independent of the weight value. Each generator is made of 2 1 × 1 convolution filters, 5 fractionally-strided convolution (FConv) filters and 9 residual blocks which can output a 224 × 224×3 (RGB) image. The stacked layers in each residual block consist of two convolutional layers where the output matrix is the same dimension as the input matrix. Each convolutional layer in the generator is followed by a batch normalization [15] and a ReLU activation except for the output layer, which uses a scaled tanh to make the output in the range [-1, 1] instead.

B. DISCRIMINATOR

The two sets of discriminators $[D_{f,k}^1, D_{s,k}^1]_{k=1,2}$ and $[D_{f,k}^2, D_{s,k}^2]_{k=1,2}$ are trained to enhance the quality of the normalized faces, where k refers to the global region and the local region. Similar discriminators have been seen in a few previous work on face frontalization [2], [5], [16]. As shown in Figure 1, the global region, enclosed by the red bounding box, covers the whole face region, including the hair, ear and some background. The local region, enclosed by the cyan bounding box, covers the face only without the background. Both global and local regions are defined for frontal- and side-view normal sets by using the facial landmarks detected by the Face Alignment Network [17].

The conventional way to discriminate the real data distribution from the generated one considers the cross-entropy loss with the Jensen-Shannon (JS) divergence [9], [18]. However, it is a common observation that the discriminator built upon minimizing the JS divergence is often hard to train and may wind up to the mode collapse, where the generator can only make limited types of data [19]. We, therefore, adopt the Wasserstein GAN with gradient penalty (WGAN-GP) loss for the discriminators [19], and it can be formulated as:

$$L_{f,a}^1 = D_{f,k}^1(\tilde{y}_f) + D_{f,k}^1(\tilde{x}_{f,f}) + D_{f,k}^1(\tilde{x}_{s,f}) - D_{f,k}^1(x_f) + \lambda_{pen}(||\nabla_{\check{x}_f} D_{f,k}^1(\check{x}_f) - 1||_2)^2, k \in [1,2] \quad (7)$$

where $L_{f,a}^1$ denotes the frontal-view adversarial loss at Layer 1. $[D_{f,k}^1]_{k=1,2}$ denotes the pair of the frontal-view discriminators, where $k = 1$ is for the global region and $k = 2$ is for the local region. λ_{pen} denotes the penalty coefficient. The last term $||\nabla_{\check{x}_f} D_{f,k}^1(\check{x}_f) - 1||_2$ is the penalty on the gradient norm computed at a random sample $\check{x}$, which is implicitly defined as the distribution of the uniform samples from the straight lines between the pairs of the frontal-view real and generated data. For more details and settings about the penalty function and WGAN-GP, please refer to [19]. Similarly, $L_{s,a}^1$, the side-view adversarial loss at Layer 1, can be expressed as follows:

$$L_{s,a}^1 = D_{s,k}^1(\tilde{y}_s) + D_{s,k}^1(\tilde{x}_{f,s}) + D_{s,k}^1(\tilde{x}_{s,s}) - D_{s,k}^1(x_s) + \lambda_{pen}(||\nabla_{\check{x}_s} D_{s,k}^1(\check{x}_s) - 1||_2)^2, k \in [1,2] \quad (8)$$

The adversarial loss also needs to be defined for Layer-2 generators. The Layer-2 frontal-view adversarial loss $L_{f,a}^2$ can be calculated as follows:

$$\begin{aligned} L_{f,a}^2 = {} & D_{f,k}^2(\hat{y}_f) + D_{f,k}^2(\hat{x}_{f,f}) + D_{f,k}^2(\hat{x}_{s,f}) \\ & + D_{f,k}^2(\hat{x}_{f,f,f}) + D_{f,k}^2(\hat{x}_{f,s,f}) + D_{f,k}^2(\hat{x}_{s,f,f}) \\ & + D_{f,k}^2(\hat{x}_{s,s,f}) - D_{f,k}^2(\hat{y}_{f,f}) + D_{f,k}^2(\hat{y}_{s,f}) - D_{f,k}^2(x_f) \\ & + \lambda_{pen}(||\nabla_{\check{x}_f} D_{f,k}^2(\check{x}_f) - 1||_2)^2, k \in [1,2] \end{aligned} \quad (9)$$

where $\hat{y}_f$, $\hat{x}_{f,f}$ and $\hat{x}_{s,f}$ denote the synthesized faces for the data from S_m, $S_{n,f}$ and $S_{n,s}$, respectively. The side-view adversarial loss $L_{s,a}^2$ at Layer-2 can be similarly defined as:

$$\begin{aligned} L_{s,a}^2 = {} & D_{s,k}^2(\hat{y}_s) + D_{s,k}^2(\hat{x}_{f,s}) + D_{s,k}^2(\hat{x}_{s,s}) \\ & + D_{s,k}^2(\hat{x}_{f,f,s}) + D_{s,k}^2(\hat{x}_{f,s,s}) + D_{s,k}^2(\hat{x}_{s,f,s}) \\ & + D_{s,k}^2(\hat{x}_{s,s,s}) - D_{s,k}^2(\hat{y}_{f,s}) + D_{s,k}^2(\hat{y}_{s,s}) - D_{s,k}^2(x_s) \\ & + \lambda_{pen}(||\nabla_{\check{x}_s} D_{s,k}^2(\check{x}_s) - 1||_2)^2, k \in [1,2] \end{aligned} \quad (10)$$

The network settings of the discriminators are shown in the right part of Table 1. The discriminators $D_{k=1}$ and $D_{k=2}$ are made of 5 and 4 convolution layers, respectively, followed by a fully connected layer to produce a single output. The images that enter $D_{k=1}$ are sized $224 \times 224 \times 3$, and the images with local regions that enter $D_{k=2}$ are sized 150 × 156 × 3. The output shows the probability of the input image being real that is measured by the Wasserstein distance. Follow the WGAN-GP [19], we use layer normalization [20] instead of batch normalization.

C. OBJECTIVE FUNCTIONS

We also consider the following loss functions when training the generators:

- **Symmetry Loss**: Although most human faces are not perfectly symmetric, symmetry is an explicit characteristic of a frontal face for the most part. We exploit this common observation as a prior and impose a symmetry loss on the generation of frontal-view faces to alleviate the self-occlusion issues cased by large poses. The symmetry loss L^1_{sym} for the Layer-1 frontal-view generator G^1_f is calculated as follows:

$$L^1_{sym} = |\tilde{y}_f - \tilde{y}'_f| + |\tilde{x}_{s,f} - \tilde{x}'_{s,f}| + |\tilde{x}_{f,f} - \tilde{x}'_{f,f}| \quad (11)$$

where $\tilde{x}'_{f,f}$, $\tilde{x}'_{s,f}$ and $\tilde{y}'_f$ denote the flip version of the frontal-view normalized faces $\tilde{x}_{f,f}$, $\tilde{x}_{s,f}$ and $\tilde{y}_f$, respectively. Another symmetry loss can be defined for the Layer-2 frontal-view generator G^2_f as follows:

$$\begin{aligned} L^2_{sym} = {} & |\tilde{y}_f - \tilde{y}'_f| + |\tilde{x}_{s,f} - \tilde{x}'_{s,f}| + |\tilde{x}_{f,f} - \tilde{x}'_{f,f}| \\ & + |\tilde{y}_{f,f} - \tilde{y}'_{f,f}| + |\tilde{y}_{s,f} - \tilde{y}'_{s,f}| \\ & + |\tilde{x}_{f,s,f} - \tilde{x}'_{f,s,f}| + |\tilde{x}_{s,f,f} - \tilde{x}'_{s,f,f}| \\ & + |\tilde{x}_{s,s,f} - \tilde{x}'_{s,s,f}| + |\tilde{x}_{f,f,f} - \tilde{x}'_{f,f,f}| \end{aligned} \quad (12)$$

- **Reconstruction Loss**: The reconstruction loss can be used to better preserve the perceptual pattern of the input. We consider the pixel-based L_1 loss between the input and output when training on the normal sets. When training G^1_f and G^1_s, we adopt the following reconstruction loss:

$$L^1_{rec} = |x_f - \tilde{x}_{f,f}| + |x_f - \tilde{x}_{s,f}| + |x_s - \tilde{x}_{f,s}| + |x_s - \tilde{x}_{s,s}| \quad (13)$$

As the training of G^2_f and G^2_s includes the Layer-1 generated data, the reconstruction loss L^2_{rec} must include the dual-view normalized faces and can be written as follows:

$$\begin{aligned} L^2_{rec} = {} & |x_f - \hat{x}_{f,f,f}| + |x_f - \hat{x}_{f,s,f}| + |x_f - \hat{x}_{s,f,f}| \\ & + |x_f - \hat{x}_{s,s,f}| + |x_s - \hat{x}_{f,f,s}| + |x_s - \hat{x}_{f,s,s}| \\ & + |x_s - \hat{x}_{s,f,s}| + |x_s - \hat{x}_{s,s,s}| + |\tilde{y}_f - \hat{y}_{f,f}| \\ & + |\tilde{y}_f - \hat{y}_{s,f}| + |\tilde{y}_s - \hat{y}_{f,s}| + |\tilde{y}_s - \hat{y}_{s,s}| \\ & + |x_f - \tilde{x}_{f,f}| + |x_f - \tilde{x}_{s,f}| \\ & + |x_s - \tilde{x}_{f,s}| + |x_s - \tilde{x}_{s,s}| \end{aligned} \quad (14)$$

The overall objective function for training the DVN can be summarized in the following composite losses:

$$\begin{cases} L^1_{G,f} = \lambda_a L^1_{f,a} + \lambda_{rec} L^1_{rec} + \lambda_{id} L^1_{id} + \lambda_{sym} L^1_{sym} \\ L^1_{G,s} = \lambda_a L^1_{s,a} + \lambda_{rec} L^1_{rec} + \lambda_{id} L^1_{id} \\ L^2_{G,f} = \lambda_a L^2_{f,a} + \lambda_{rec} L^2_{rec} + \lambda_{id} L^2_{id} + \lambda_{sym} L^2_{sym} \\ L^2_{G,s} = \lambda_a L^2_{s,a} + \lambda_{rec} L^2_{rec} + \lambda_{id} L^2_{id} \end{cases} \quad (15)$$

D. THREE-PHASE TRAINING

To better train the DVN framework, we adopt a three-phase training scheme. In the first phase, we train the Layer-1 generators G^1_f and G^1_s by using the normal sets and source set until the performance reaches a steady level, i.e., the rank-1 identification rate reaches 95% within 45° of yaw on Multi-PIE and 88% on IJB-A. In the second phase, we train the Layer-2 generators G^2_f and G^2_s by using the trained $G^{1,*}_f$ and $G^{1,*}_s$ as the initial models with the generated data $[\tilde{x}, \tilde{y}]$, the normal sets, and source set as the training set. The training continues until the performance reaches some steady level, i.e., the rank-1 identification rate reaches 97% within 45° of yaw on Multi-PIE and 90% on IJB-A. The last phase concatenates Layer-1 and Layer-2 and allows $G^{1,*}_f$, $G^{1,*}_s$, $G^{2,*}_f$ and $G^{2,*}_s$ to interact with each other within the DVN framework to further improve the overall performance.

IV. EXPERIMENTAL EVALUATION

In the following, we first present the experimental settings and protocols in Sec. IV-A, then an ablation study for determining the model parameters and settings in Sec. IV-B, and then the performance comparison with other state-of-the-art methods in Sec. IV-C.

A. SETTINGS AND PROTOCOLS

Both of the constrained and unconstrained settings were considered in our experiments. The constrained setting was experimented on the Multi-PIE dataset [21], and the unconstrained setting was experimented on the IJB-A [22] and IJB-C [23] datasets. For the constrained settings, the Multi-PIE was split into a training set and a testing set. The former was used to train the DVN framework, and the latter was used to evaluate the performance of using the DVN-generated images for face recognition. For the unconstrained settings, the DVN was trained on the combination of the Multi-PIE and CASIA-WebFace, and the face recognition performance was evaluated on the IJB-A and IJB-C datasets.

1) CONSTRAINED SETTINGS

The Multi-PIE [21] is one of the most popular in-the-house databases, and it contains more than 750,000 images of 337 people recorded in four sessions over the span of five months. The subjects were imaged under 15 view points and 19 illumination conditions with a range of facial expressions. To demonstrate how DVN maintains the identity of input faces, we used the face images of the 250 subjects in Session-1 for the experiments on the constrained settings. The first 150 subjects were chosen for training and further divided into two normal sets and a source set. Both normal sets, i.e., the frontal- and side-view normal sets, contain 750 face images in neural expression with 5 illuminations conditions and the rest 37,500 face images form the source set. The frontal- and side-view sets were chosen from two poses, labeled in 05_1 and 19_0, under the illumination conditions with Lighting-labels 05∼09 and 06∼10, respectively. For performance testing, the frontal-view pose with evenly distributed

illumination (Lighting 07) and neutral expression was chosen as the gallery image for each of the remaining 100 people, and the rest images were used as probes. Similar recognition protocol can be found in previous work, e.g., [2], [8] [5].

To demonstrate the recognition performance, we followed the generation-for-recognition protocol that all probe images are first normalized and matched against the gallery images. For the constrained setting on the Multi-PIE, the gallery images are by nature normalized, and only the non-frontal faces are normalized. We employed both ArcFce [1] and Light-CNN [24] as feature extractor for distilling the identity feature of each normalized face, and compared the cosine distance between the probe and gallery features. For the unconstrained setting on the IJB-A database, both the gallery images and probe images are normalized, then the features are extracted by the feature extractor, and then the cosine distances are computed.

For pre-processing, each image was first processed by the Face Alignment Network [17] for face and facial landmark detection. All faces were aligned to the eyes, and normalized in scale to make the distance between the mouth and the center of both eyes 86 pixels. Each normalized image is 250×250 in scale. The DVN training was proceeded via the Adam optimizer [25] with initial learning rate 0.001, momentum 0.5, and batch size 16. The implementation was on the Tensorflow [26] using a two NVIDIA TITAN RTX GPU with 48G memory.

2) UNCONSTRAINED SETTINGS

For the unconstrained settings, we trained the DVN network on the Multi-PIE and CASIA-WebFace [27], and evaluated face recognition on the IJB-A [22] and IJB-C [23]. The normal sets were formed by the frontal- and side-view images with the same pose and illumination labels as those selected in the above constrained settings, but taken from all 337 subjects of the entire Multi-PIE. The CASIA-WebFace offers 494,414 images of 10,575 people taken *in the wild*. The images in the CASIA-WebFace form the source set. Due to the label noise in the original database, we cleaned it and extracted 455,577 face images.

The IJB-A offers 5,396 images and 20,412 video frames of 500 subjects. It defines template-to-template test protocol and each template has one or more images. The evaluation protocol in IJB-A consists of template-to-template verification and identification over 10 splits. Each split contains around 11,748 pairs (1,756 positive and 9,992 negative pairs) for verification and 112 gallery templates and 1,763 probe templates for identification on average. Moreover, some templates contain only one challenging image, i.e., large pose and low image quality. These factors essentially make IJB-A challenging face recognition dataset. The IJB-C is an extension of IJB-A, which contains 3,134 still images and 117,542 frames from natural scene video of 3,531 different individuals. It is designed to have a more uniform geographic distribution of subjects across globe, which make it possible to carefully evaluate many covariates in details. The IJB-C 1:1 mixed verification protocol provides 19,557 genuine pairs and 15,638,932 imposter pairs for evaluations of performance. The preprocessing was the same as that for the constrained settings on the Multi-PIE.

The verification and identification were performed based on those normalized gallery and probe images. This is slightly different than that in the constrained case where the gallery images are all frontal, evenly lit and in neural expression, i.e., already normalized by their nature.

B. ABLATION STUDY

As the characteristics of the proposed DVN include 1) the generation of two normalized views, 2) the double layers of dual-view generators, the ablation study was designed for the following inspections:

1) Due to the fact that the Multi-PIE offers the ground-truth of cross-pose faces for the same subjects, the experiments under the constrained settings allow the comparison of the ground-truth with the DVN generated dual-view faces, which can further be used for evaluating the generation-for-recognition performance.
2) To better compose the DVN, we have to determine the best weights λ_a, λ_{id}, λ_{rec}, λ_{sym} in the loss function (15) through experiments. In all our experiments, we use the penalty coefficient $\lambda_{pen} = 10$ in (7), (8), (9) and (10), which is empirically validated in [19]. To shorten the time consumption, the experiments were carried out on the single-view normalization network, which only generated one of the three views (frontal, side and profile), under the unconstrained settings. The results in Table 3 denote the single frontal view as "SF", the single side view as "SS", and we also added in a single profile view (SP) solely for the comparison purpose as forensics often collect frontal and profile views. Note that for the experiment in the constrained settings, we ran a similar experiment on the training dataset for determining the weights.
3) As the DVN can generate two normalized views, we ran experiments to compare the performance of using only one single-view normalized faces generated by the DVN and that by using the single-view normalization network mentioned in the above 2). The former is denoted as SF (DVN-S) and SS (DVN-S) in Table 3 when using Layer-1 generators only (see the following item 4 for more details).
4) As the DVN is composed of double layers of generators, the experiments were designed to compare the performance of using Layer-1 generators only and that of using both layers of generators. The former is denoted as DVN-S for single layer, and the latter as DVN-D for double layers.
5) Two schemes were attempted to fuse the features of the dual-view normalized faces, one is the average of the dual-view latent vectors, denoted as DVN-S_m, DVN-D_m, and the other is the concatenation of the

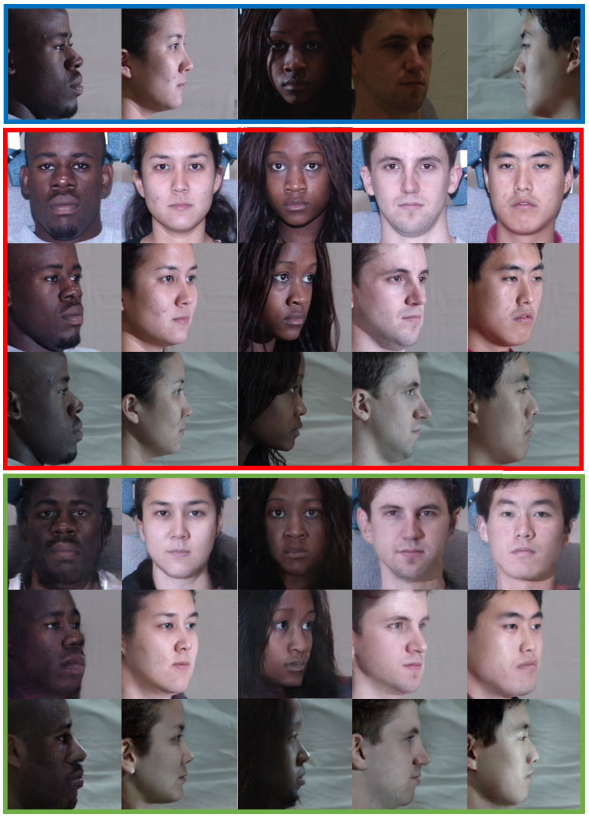

FIGURE 2. DVN generated faces compared with the ground truth. Top row shows the input faces. Row-2 to 4, enclosed by the red bbox, are the ground-truth in frontal, 45° side and profile poses, respectively. Row-5 to 7, enclosed by the green bbox, are the generated faces in corresponding poses.

dual-view latent vectors, denoted as DVN-S_c and DVN-D_c.

The comparison of the DVN generated face images with the ground-truth of several subjects randomly selected from the Multi-PIE is shown in Figure 2. The generated normalized images preserve the identity characteristics of each face to some extent, and the frontal-views appear to preserve better than the side- and profile-views. Table 2 shows the Rank-1 cross-pose recognition rates. The DVN (Light-CNN) and DVN (ArcFace) refer to using the DVN-normalized face images as input, and the Light-CNN [28] and ArcFace [1] respectively as the feature extractor. For comparison purpose, we also include the performance of the original Light-CNN and ArcFace with the un-normalized faces as inputs. Note that the encoder E_n in our framework is made of the ArcFace, using the ArcFace as the feature extractor will lead to better performance, which is also shown in Table 2.

TABLE 2. Performance comparison on Multi-PIE.

Method	15°	30°	45°	60°	75°	90°
FF-GAN [6]	94.6	92.5	89.7	85.2	77.2	61.2
TP-GAN [2]	99.8	99.9	98.6	98.1	92.9	75.0
DR-GAN [10]	95.0	91.3	88.0	85.8	-	-
Light-CNN [28]	99.2	98.0	97.7	95.5	73.3	20.7
CAPG-GAN [8]	99.9	99.4	98.3	93.7	87.4	77.1
PIM [3]	99.3	99.0	98.5	98.1	95.0	86.5
HF-PIM [4]	99.9	99.9	99.9	98.1	96.4	92.3
ArcFace [1]	99.9	99.9	99.9	99.1	79.6	40.2
FNM [5]	99.9	99.5	98.2	93.7	81.3	55.8
DVN (Light-CNN)	99.9	99.9	99.8	99.0	95.8	85.1
DVN (ArcFace)	99.9	99.9	99.9	99.4	96.9	88.6

While the performance of many other methods degrades for extreme pose, i.e., $\geq 75°$ in yaw, the DVN outperforms others for almost all viewing angles, except for 90°, where the DVN is the second best, outperformed by the HF-PIM [4]. The HF-PIM requires paired data for learning and may capture the intrinsic characteristics between the frontal and profile poses. Such pair-wise characteristics will be considered in the improvement phase of the DVN.

The experiments for the determination of the best weights are shown in the top part of Table 3, where all experiments were conducted on the single-view normalization network under unconstrained settings and the Light-CNN was exploited as the feature extractor. The weight for the identity loss λ_{id} changes from 0 (i.e., no identity loss) to 5500 while other weights are kept as follows: $\lambda_a = 1$, $\lambda_{sym} = 0.01$ and $\lambda_{rec} = 0.001$; and for cases when λ_a, λ_{sym} and λ_{rec} are switched to zero, i.e., excluding these losses. The aforementioned weights for λ_a, λ_{sym} and λ_{rec} were determined through similar experiments; however, the influences of these three weights were not as strong as λ_{id}. We show 4 cases of λ_{id}, 0, 1000, 3500 and 5500 in Table 3. Figure 3 shows the corresponding normalized faces for visual comparison.

TABLE 3. Quantitative results of ablation study. All experiments exploit the ArcFace as the feature extractor.

Model	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
SF ($\lambda_{id} = 1000$)	87.0±0.4	77.7±1.0	92.2±1.7	94.8±0.9
SF ($\lambda_{id} = 3500$)	91.0±0.8	81.7±0.9	93.5±1.9	97.3±1.0
SF ($\lambda_{id} = 5500$)	89.5±0.9	77.1±2.9	91.0±1.2	95.8±0.8
SF ($\lambda_{id} = 0$)	7.5±1.5	2.8±1.2	9.3±1.6	25.7±2.1
SF ($\lambda_a = 0$)	85.3±0.7	68.1±1.9	83.6±1.2	91.4±1.4
SF ($\lambda_{sym} = 0$)	86.6±1.4	76.3±1.3	86.6±1.6	92.3±1.2
SF ($\lambda_{rec} = 0$)	88.0±1.1	79.3±1.0	88.6±2.3	93.9±1.7
SS	87.0±1.1	79.2±0.7	90.5±1.9	94.6±0.7
SP	54.5±0.9	32.5±0.7	77.9±0.7	85.2±1.3
SF (DVN-S)	91.4±0.3	82.2±1.7	93.9±1.2	97.4±0.9
SS (DVN-S)	87.9±0.6	80.9±1.2	91.2±1.1	95.2±1.9
SF (DVN-D)	92.2±0.8	84.0±1.9	94.2±0.7	97.6±0.9
SS (DVN-D)	90.0±0.6	81.4±0.6	92.3±0.6	96.1±1.2
DVN-S_c	93.2±0.8	85.0±1.0	94.8±1.1	97.2±0.9
DVN-S_m	94.1±1.8	86.2±1.9	95.8±0.1	98.1±0.9
DVN-D_c	94.2±0.5	86.3±0.2	95.7±0.9	98.0±0.9
DVN-D_m	97.2±1.1	94.2±0.6	97.4±0.5	98.8±0.2

The faces generated with $\lambda_{id} = 0$ are almost impossible to recognize as those faces show a clear collapsed mode. The performance improves substantially with increasing λ_{id}, and reaches the best when $\lambda_{id} = 3500$. In the case without the

FIGURE 3. Face normalization on IJB-A for different weight setups, including, (a) $\lambda_{id} = 1000$, (b) $\lambda_{id} = 3500$, (c) $\lambda_{id} = 5500$, (d) $\lambda_{id} = 0$, (e) $\lambda_{a} = 0$, (f) $\lambda_{sym} = 0$, (g) $\lambda_{rec} = 0$. Based on the optimal setup above, we also show the normalized faces in multiple-view, including (g) Side-view, and (h) Profile-view.

adversarial loss ($\lambda_a = 0$), the generated faces demonstrate strong artifacts and collapsed mode. The poor quality of the generated faces are also revealed by the low recognition rates in Table 3. The contributions of the symmetry loss and reconstruction loss are marginal, and the former is slightly more important than the latter, as shown in the table.

The performance of SF (Single-View Frontal, $\lambda_{id} = 3500$), SS (Single-View Side) and SP (Single-View Profile) shows that the frontal view performs the best, then the side view, and then the profile view. The performance gaps can be partially due to the imbalanced pose distribution of the test dataset which contains far less data with extreme poses than those with the frontal and frontal-to-side poses. As the SP degrades the recognition performance, we exclude it from the desired dual views.

The middle part of Table 3 shows the performance of SF (DVN-S), SS (DVN-S), SF (DVN-D) and SS (DVN-D). The DVN architecture is verified capable of improving the performance of the single-view normalization on both frontal- and side-views, even by using only Layer-1 generators. It also demonstrates that the double-layer generators lead to better performance than the single-layer generators

As far as the feature fusion scheme is concerned, the mean feature, i.e., the average of the latent vectors of the frontal- and side-views performs better than the concatenated one. This is demonstrated in the bottom part of Table 3 by the performances of both the single-layer generators (in DVN-S_c and DVN-S_m), and the double-layer generators (in DVN-D_c and DVN-D_m). Several dual-view normalized faces with the original inputs from the IJB-A are shown in Figure 4. The normalized faces made by the Layer-1 generators appears slightly worse than those made by the Layer-2 generators, especially the side-view faces

TABLE 4. Performance comparison on IJB-A.

Method	Verification		Identification	
Metric(%)	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
FF-GAN [6]	85.2±1.0	66.3±3.3	90.2±0.6	95.4±0.5
DR-GAN [10]	87.2±1.4	78.1±3.5	92.0±1.3	96.1±0.7
Light-CNN [28]	91.2±1.1	84.4±0.8	92.4±1.7	95.4±0.8
HF-PIM [4]	95.2±0.7	89.7±1.4	96.1±0.5	97.9±0.2
ArcFace [1]	96.6±1.2	93.5±0.5	96.7±0.6	98.1±0.3
FNM [5]	93.4±0.9	83.8±2.6	96.0±0.5	98.6±0.3
DNV (Light-CNN)	95.7±0.5	91.3±1.7	96.8±1.0	98.7±0.4
DVN (ArcFace)	97.2±1.1	94.2±0.6	97.4±0.5	98.8±0.2

TABLE 5. Performance comparison on IJB-C.

	Verification	
Method	@FAR=.01	@FAR=.001
FaceNet [29]	32.40	20.58
VGGFace [30]	45.60	26.18
DR-GAN [10]	88.20	73.60
VGGFace2 [31]	95.00	90.00
Light-CNN [28]	90.63	84.32
FNM [5]	89.82	80.23
ArcFace [1]	95.82	91.69
Proposed (Light-CNN)	92.43	87.96
Proposed (ArcFace)	96.55	92.76

C. COMPARISON WITH STATE-OF-THE-ART APPROACHES

Table 4 shows the verification and identification rates on the IJB-A with a comparison to state-the-art approaches. As it is verified in the above ablation study that the double-layered DVN-D outperforms the single-layered DVN-S, we only report the comparison with the DVN-D and simply write it as DVN. For comparison purpose, we also include the performance of using the Light-CNN as the feature extractor, and denote it as DVN (Light-CNN). The DVN (ArcFace) outperforms all selected methods for both verification and identification, and the performance gaps for the verification rate at FAR 0.001 and Rank-1 identification rate are clear to observe. Furthermore, as expected, the DVN (ArcFace) performs better than DVN (Light-CNN) since the face encoder in the DVN is made of the ArcFace.

FIGURE 4. The frontal- and side-view synthesized faces on the IJB-A. The top row shows the input faces. Rows 2 and 3, labeled by (a) and enclosed by a red bounding box, are the normalized dual-view faces made by Layer-1 generators. Rows 4 and 5, labeled by (b) and enclosed by a green bounding box, are the normalized dual-view faces generated by Layer-2 generators.

Table 5 shows the performance evaluation on the IJB-C verification protocol with other state-of-the-art approaches. The IJB-C is generally considered more challenging than the IJB-A. However, similar results are obtained, as shown in the table. The DVN (ArcFace) outperforms all, and performs better than the DVN (Light-CNN). The comparisons verify the effectiveness of the proposed DVN.

V. CONCLUSION

Frontal face is generally considered as the only standard for face normalization. It is, however, also commonly acknowledged that a face can be better characterized by multiple views, as the case in forensics, where both frontal and profile poses are kept. The proposed DVN (Dual-View Normalization) can be the first that proposes face normalization in dual poses, and verifies the effectiveness of the dual-view normalization with a specially designed double-layer architecture. The design of the DVN consists of identity preservation, normalized pose transformation and (source-target) domain transformation. Experiments show that the DVN outperforms many state-of-the-art approaches for face normalization, and can lead to better performance for face recognition. As for the potential extension of this research, we consider the unsupervised clustering of facial attributes for preprocessing the constrained and unconstrained datasets, so that the learning for normalization can be made attribute-oriented. The approach proposed in [32] can be a decent example.

REFERENCES

[1] 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.
[2] 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.
[3] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. CVPR*, Jun. 2018, pp. 2207–2216.
[4] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. NIPS*, 2018, pp. 2867–2877.
[5] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9851–9858.
[6] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3990–3999.
[7] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. SIGGRAPH*, 1999, pp. 187–194.
[8] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.
[9] 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.
[10] L. Q. 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.
[11] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.
[12] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: An approach for best exemplar selection," *IEEE Trans. Image Process.*, vol. 29, pp. 5457–5468, 2020.

[13] H. Yuan, M. Li, J. Hou, and J. Xiao, "Single image-based head pose estimation with spherical parametrization and 3D morphing," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107316.
[14] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "LFR face dataset: Left-front-right dataset for pose-invariant face recognition in the wild," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Feb. 2020, pp. 124–130.
[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[16] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Proc. NIPS*, 2017, pp. 66–76.
[17] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. NIPS*, 2017, pp. 5767–5777.
[20] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.
[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.
[23] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA janus Benchmark–C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 158–165.
[24] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation," *Comput. Res. Repository*, 2015, *arXiv: 1511.02683*. [Online]. Available: https://arxiv.org/abs/1511.02683
[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[26] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. OSDI*, 2016, pp. 265–283.
[27] 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
[28] 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.
[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[30] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4362–4371.
[31] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[32] A. F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, and F. Narducci, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access*, vol. 8, pp. 9037–9045, 2020.

GEE-SERN (JISON) HSU (Senior Member, IEEE) received the dual M.S. degrees in electrical and mechanical engineering and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, in 1993 and 1995, respectively. From 1995 to 1996, he was a Postdoctoral Fellow with the University of Michigan. From 1997 to 2000, he was a Senior Research Staff with the National University of Singapore. In 2001, he joined PenPower Technology, where he led research on face recognition and intelligent video surveillance. In 2007, he joined the Department of Mechanical Engineering, National Taiwan University of Science and Technology (NTUST), where he is currently an Associate Professor. His research interests include computer vision and pattern recognition. He is a Senior Member of IAPR. He received the Best Paper Awards in ICMT 2011, CVGIP 2013, CVPRW 2014, ARIS 2017, and CVGIP 2018. His team at PenPower Technology was a recipient of the Best Innovation and the Best Product Award at SecuTech Expo, for 3 consecutive years, from 2005 to 2007.

CHIA-HAO TANG (Student Member, IEEE) received the B.S. degree in mechanical engineering from Yuan Ze University, Taoyuan City, Taiwan, in 2017, and the M.S. degree in mechanical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include deep learning and computer vision, in particular, face recognition and generative adversarial networks.