3D-Aided Dual-Agent GANs for Unconstrained Face Recognition

Jian Zhao, *Student Member, IEEE,* Lin Xiong, Jianshu Li, Junliang Xing, *Senior Member, IEEE*, Shuicheng Yan, *Fellow, IEEE*, and Jiashi Feng, *Member, IEEE*

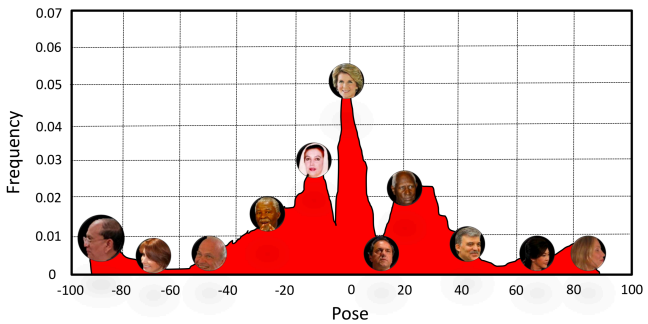
Abstract—Synthesizing realistic profile faces is beneficial for more efficiently training deep pose-invariant models for large-scale unconstrained face recognition, by augmenting the number of samples with extreme poses and avoiding costly annotation work. However, learning from synthetic faces may not achieve the desired performance due to the discrepancy between distributions of the synthetic and real face images. To narrow this gap, we propose a **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) model, which can improve the realism of a face simulator's output using *unlabeled* real faces while preserving the identity information during the realism refinement. The dual agents are specially designed for distinguishing real *v.s.* fake and identities simultaneously. In particular, we employ an off-the-shelf 3D face model as a simulator to generate profile face images with varying poses. DA-GAN leverages a fully convolutional network as the generator to generate high-resolution images and an auto-encoder as the discriminator with the dual agents. Besides the novel architecture, we make several key modifications to the standard GAN to preserve pose, texture as well as identity, and stabilize the training process: (i) a pose perception loss; (ii) an identity perception loss; (iii) an adversarial loss with a boundary equilibrium regularization term. Experimental results show that DA-GAN not only achieves outstanding perceptual results but also significantly outperforms state-of-the-arts on the large-scale and challenging NIST IJB-A and CFP unconstrained face recognition benchmarks. In addition, the proposed DA-GAN is also a promising new approach for solving generic transfer learning problems more effectively. DA-GAN is the foundation of our winning entry to the NIST IJB-A face recognition competition in which we secured the 1^{st} places on the tracks of verification and identification.

Index Terms—Face Synthesis, Unconstrained Face Recognition, 3D Face Model, Generative Adversarial Networks.

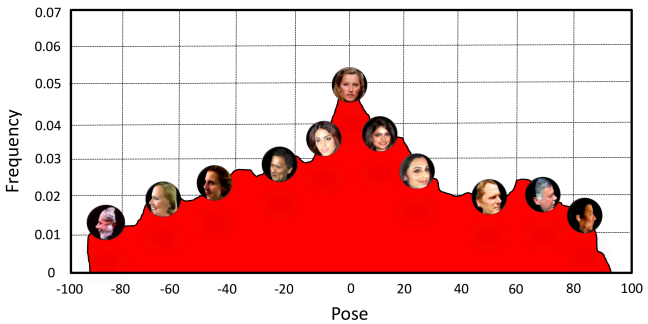
1 Introduction

Unconstrained face recognition is a very important yet challenging problem. In recent years, deep learning techniques have significantly advanced large-scale unconstrained face recognition [1], [2], [3], [4], [5], [6], arguably driven by rapidly increasing resources of face images. However, labeling huge amount of data for feeding supervised deep learning algorithms is expensive and time-consuming. Moreover, as often observed in real-world scenarios, the pose distribution of available face recognition datasets (*e.g.*, IJB-A [7]) is usually unbalanced, showing a long tail with large pose variations as in Fig. 1 (a). This has become a main obstacle for further pushing unconstrained face recognition performance. Several research attempts [8], [9], [10] have been made to employ synthetic profile face images as augmented extra data to balance the pose variations.

However, naively learning from synthetic images may be problematic due to the distribution discrepancy between synthetic and real face images—synthetic data are often not realistic enough with artifacts and severe texture loss. The low-quality synthesized face images would mislead the learned face recognition model to overfit to fake information only contained in synthetic images and fail to generalize well on real faces. Increasing the realism of the simulator by brute force is often expensive in terms of time cost and manpower, if possible.



(a) Extremely unbalanced pose distribution.



(b) Well balanced pose distribution with DA-GAN.

Fig. 1: Comparison of pose distribution in the IJB-A [7] dataset w/o and w/ DA-GAN.

- *Jian Zhao and Lin Xiong make equal contributions. Jian Zhao was an intern at Panasonic R&D Center Singapore during this work. Jian Zhao is the corresponding author. Homepage: https://zhaoj9014.github.io/.*
- *Jian Zhao, Shuicheng Yan, and Jiashi Feng are with Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Jian Zhao is also with School of Computer, National University of Defense Technology, China. Shuicheng Yan is also with Qihoo 360 AI Institute, China. E-mail: {zhaojian90, jianshu}@u.nus.edu, {elefjia, eleyans}@nus.edu.sg.*
- *Lin Xiong is with Core Technology Group, Learning & Vision, Panasonic R&D Center Singapore, Singapore. E-mail: lin.xiong@sg.panasonic.com.*
- *Junliang Xing is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. E-mail: jlxing@nlpr.ia.ac.cn.*

In this work, we propose a novel **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) for profile view synthesis, where the dual agents are responsible for discriminating the realism of synthetic profile face images generated from a simulator using unlabled real data and perceiving the identity information, respectively. In other words, the generator needs to play against a real-fake discriminator as well as an identity discriminator simultaneously to generate high-quality faces that are really useful for unconstrained face recognition.

In our method, a synthetic profile face image with a prespecified pose is generated by a 3D morphable face simulator. DA-GAN takes this synthetic face image as input and refines it through a conditioned generative model. We leverage a **F**ully **C**onvolutional **N**etwork (FCN) [11] that operates on the pixel level as the generator to generate high-resolution face images and an auto-encoder network as the discriminator. Different from vanilla GANs [12], DA-GAN introduces an auxiliary discriminative agent to enforce the generator to preserve identity information of the generated faces, which is crucial for face recognition. In addition, DA-GAN also imposes a pose perception loss to preserve pose and texture. The refined synthetic profile face images show photorealistic quality with well preserved identity information, which are used as augmented data together with real face images for pose-invariant feature learning. For stabilizing the training process of such a dual-agent GAN model, we impose a boundary equilibrium regularization term.

Experimental results show that DA-GAN achieves outstanding perceptual performance. It significantly outperforms state-of-the-arts on the large-scale and challenging **N**ational **I**nstitute of **S**tandards and **T**echnology (NIST) **I**ARPA **J**anus **B**enchmark **A** (IJB-A) [7] and **C**elebrities in **F**rontal-**P**rofile (CFP) [13] unconstrained face recognition benchmarks. With DA-GAN, we won the 1^{st} places on verification and identification tracks in the NIST IJB-A face recognition competition, also strongly proving that our "recognition via generation" framework is effective and generic. We expect it to further benefit more face recognition and transfer learning applications in the real world.

A preliminary version of this work was published in **N**eural **I**nformation **P**rocessing **S**ystems (NIPS) 2017 [14]. We extend it in terms of three aspects: 1) We propose an enforced cross-entropy loss to improve the original identity-preserving loss by reallocating hard samples into correct and safe decision area, which reduces the intra-class distance while increasing the inter-class distance. 2) We add sufficient details on parameter setting, network architecture and training procedure for reimplementation. 3) We add the experiments to reveal how DA-GAN works, including comprehensive feature space analysis, more qualitative and quantitative analyzes, face recognition best/worst case study, *etc*.

Our contributions are summarized as follows.

- We propose a novel **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) for photorealistic and identity-preserving profile face synthesis even under extreme poses.
- The proposed dual-agent architecture effectively combines prior knowledge from data distribution (adversarial training) and domain knowledge of faces (pose and identity perception losses) to exactly recover the information lost inherently in projecting a 3D face to the 2D image space.
- We present qualitative and quantitative experiments showing the possibility of a "recognition via generation" framework and achieve top performance on the challenging NIST IJB-A [7] and CFP [13] unconstrained face recognition benchmarks without extra human annotation by training deep neural networks on the refined face images together with real images. To our best knowledge, the proposed DA-GAN is the first model that automatically generates augmented data effectively for face recognition in challenging conditions and indeed improves performance. DA-GAN won the 1^{st} places on verification and identification tracks in the NIST IJB-A face recognition competition.

2 Related Work

Traditional methods address unconstrained face recognition through 2D/3D local texture warping [18], [19], [20], [21], statistical modeling [22], [23], and deep learning [2], [24], [25], [26], [27]. For instance, Hassner *et al.* [18] used a single and unmodified 3D surface to approximate the shape of all the input faces, which is shown effective for face frontalization but suffers big performance drop for profile and near-profile[1] faces due to severe texture loss and artifacts. Sagonas *et al.* [22] proposed to perform joint frontal view reconstruction and landmark detection by solving a constrained low-rank minimization problem. Yim *et al.* [27] proposed to perform face rotation from an arbitrary pose using multi-task learning. Masi *et al.* [2] proposed to tackle pose variation by using multiple pose specific models and rendered face images with the aid of 3D modelling.

As one of the most important advancements in deep generative models [28], [29], GAN has drawn substantial attention from the deep learning and computer vision community ever since it was first presented by Goodfellow *et al.* [12]. The GAN framework learns a generator network and a discriminator network with competing loss. This min-max two-player game provides a simple yet powerful way to estimate target distribution and to generate novel image samples. Mirza and Osindero [30] proposed the conditional version of GAN, conditioned on both generator and discriminator for effective image tagging. Berthelot *et al.* [31] proposed a new **B**oundary **E**quilibrium **GAN** (BE-GAN) framework paired with a loss derived from the Wasserstein distance for training GAN, which controls the trade-off between image diversity and visual quality. The success of these works motivates us to develop profile view synthesis methods based on GAN. However, the generators of previous methods usually generate images based on a random noise vector or conditioned data and the discriminator only has a single agent to distinguish real *v.s.* fake. Thus, in contrast to our method, their generated images do not have any discriminative information that can be used for training a deep learning based recognition model. This separates us well with previous GAN-based attempts.

Moreover, different from previous InfoGAN [32] which does not have the classification agent, and **A**uxiliary **C**lassifier **GAN** (AC-GAN) [33] which only performs classification, our proposed DA-GAN performs face verification with intrigued data augmentation. DA-GAN generates data in a completely different way from InfoGAN [32] and AC-GAN [33]. These traditional Gan-like models generate images from a random noisy input or abstract semantic labels, thus cannot exploit useful and rich prior information (*e.g.*, shape, pose of faces) for effective data generation and augmentation, therefore inferior to our model. They cannot fully

1. Faces with yaw angles greater than $60°$.

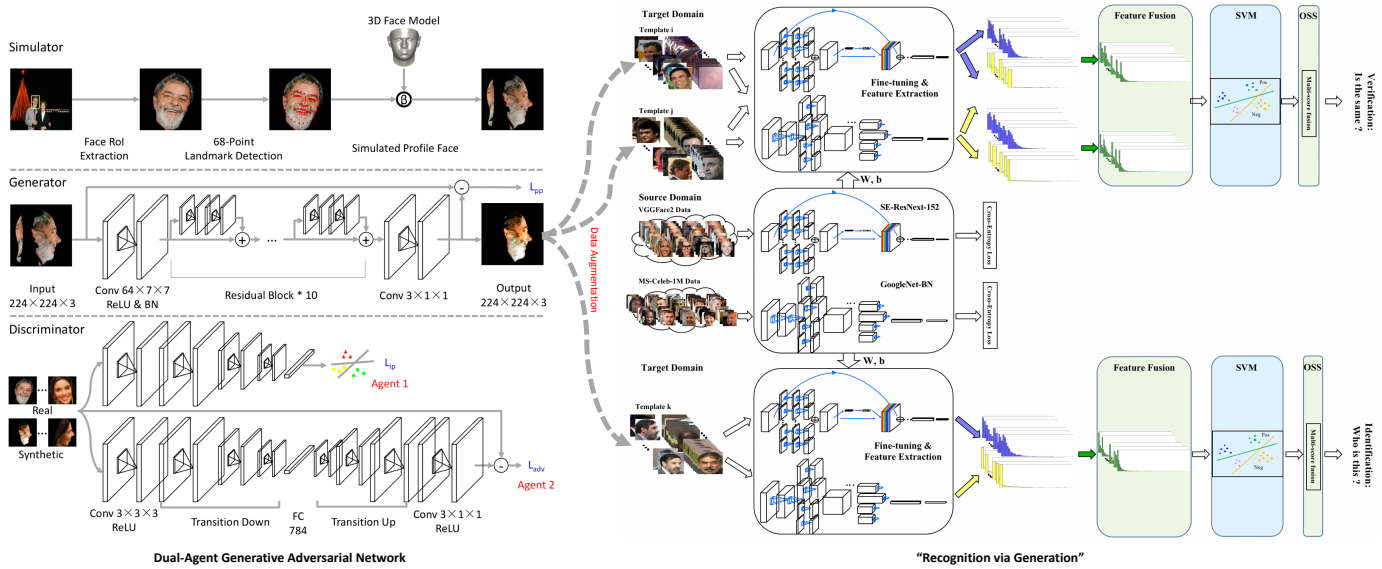


Fig. 2: Overview of the proposed unconstrained face recognition pipeline. Left panel: The proposed **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) architecture. The simulator extracts face RoI, performs saliency prediction (*i.e.*, face/background segmentation), localizes landmark points and produces synthetic faces with arbitrary poses, which are fed to DA-GAN for realism refinement. DA-GAN uses a fully convolutional skip-net as the generator and an auto-encoder as the discriminator. The dual agents are responsible for discriminating real *v.s.* fake (minimizing the loss $\mathcal{L}_{\text{adv}}$) and preserving identity information (minimizing the loss $\mathcal{L}_{\text{ip}}$). Right panel: The proposed "recognition via generation" framework. We transfer-learn two state-of-the-art deep neural networks, SE-ResNeXt-152 [15], [16] and GoogleNet-BN [17], from source domain to target domain extended by DA-GAN. We combine the complementary two-view information from the two models to train template-adapted **S**upport **V**ector **M**achine**s** (SVMs) [1]. The resulted margins are robust and discriminative for unconstrained face recognition. Best viewed in color.

control the generated images, while our DA-GAN can. With our DA-GAN, the usually very unbalanced face pose (*e.g.*, yaw angles) distribution in the real world can be adjusted. Therefore, DA-GAN can facilitate training more accurate face analysis models to solve the large pose variation problem and other relevant problems in unconstrained face recognition.

Our proposed DA-GAN shares a similar idea with **T**wo-**P**athway **GAN** (TP-GAN) [24], **F**ace **F**rontalization **GAN** (FF-GAN) [26] and **D**isentangled **R**epresentation learning **GAN** (DR-GAN) [25] that address face synthesis based on the GAN framework, and Apple GAN [34] that learns from simulated and unsupervised images through adversarial training. Our method differs from them in following aspects: 1) DA-GAN aims to synthesize photorealistic and identity-preserving profile faces to address the large variance issue in unconstrained face recognition, while TP-GAN [24], FF-GAN [26] and DR-GAN [25] try to recover a frontal face from a profile view and Apple GAN [34] is designed for much simpler scenarios (*e.g.*, eye and hand image refinement); 2) TP-GAN [24], FF-GAN [26], DR-GAN [25] and Apple GAN [34] suffer from categorical information loss which limits their effectiveness in recognition. In contrast, our proposed DA-GAN architecture effectively overcomes this issue by introducing dual discriminator agents. Experimental comparisons are provided in Sec. 4.

3 Dual-Agent GAN

3.1 Simulator

The main challenge for unconstrained face recognition lies in the large variation and few profile face images for each subject, which hinders the learning of a well-performing pose-invariant model. To address this problem, we simulate face images with various pre-defined poses (*i.e.*, yaw angles), which explicitly augments the available training data without extra human annotation efforts and balances the pose distribution. In particular, as shown in Fig. 2, we first extract the face **R**egion **o**f **I**nterest (RoI) from each available real face image, perform saliency prediction (*i.e.*, face/background segmentation), and estimate 68 facial landmark points using the **R**ecurrent **A**ttentive-**R**efinement (RAR) [9] + **i**ntegrated **F**ace **A**nalytics **N**etwork (iFAN) [35] framework, which is robust to illumination changes and does not require a shape model in advance.

The saliency prediction is incorporated based on the following considerations: 1) Since we focus on profile face synthesis (*i.e.*, profile face simulation with **3D M**orphable **M**odel (3D MM) and realism refinement with DA-GAN), the background is not the element of interest. Moreover, there may be background distortion during the profile face simulation via 3D MM. Removing background ensures DA-GAN to focus more on enhancing face realism, with reduced training difficulty. 2) The deep face recognition models are pre-trained on the large-scale database that covers multiple modalities. Therefore, the models have robustness against background variance.

We then estimate a transformation matrix between the detected 2D landmarks and the corresponding landmarks in the 3D MM using least-squares fit [10]. Finally, we simulate profile face images in various poses with pre-defined yaw angles.

However, the performance of the simulator decreases dramatically under large poses (*e.g.*, yaw angles $\in \{[-90°, -60°] \cup [+60°, +90°]\}$) due to artifacts and severe texture loss, misleading the network to overfit to fake information only contained in synthetic images and fail to generalize well on real data.

3.2 Generator

In order to generate photorealistic and identity-preserving profile-view face images which are truly beneficial for unconstrained face recognition, we further refine the above-mentioned simulated profile face images with the proposed DA-GAN.

Inspired by the recent success of FCN-based methods on image-to-image applications [11], [36] and the leading performance of skip-net on recognition tasks [15], [37], we modify a skip-net (ResNet [37]) to an FCN-based architecture as the generator $G_\theta : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{H\times W\times C}$ of DA-GAN to learn a highly non-linear transformation for profile face image refinement, where θ denotes the network parameters for the generator, and H, W, and C denote the image height, width and channel number, receptively.

Contextual information from global and local regions compensates each other and naturally benefits face recognition. The hierarchical features within a skip-net are multi-scale in nature due to the increasing receptive field sizes, which are combined together via skip connections. Such a combined representation comprehensively preserves the contextual information, which is crucial for artifact removal, fragment stitching and texture padding. Moreover, the FCN-based architecture is advantageous for generating high-resolution image-level results. More details are provided in Sec. 4.

Formally, let the simulated profile face image be denoted by x and the refined face image be denoted by $\tilde{x}$. Then

$$\tilde{x} := G_\theta(x). \tag{1}$$

The key requirements for DA-GAN are that the refined face image $\tilde{x}$ should look like a real face image in appearance while preserving the intrinsic identity and pose information from the simulator.

To this end, we propose to learn θ by minimizing a combination of three losses:

$$\mathcal{L}_{\mathrm{G}_\theta} = (-\mathcal{L}_{\mathrm{adv}} + \lambda_1\mathcal{L}_{\mathrm{ip}}) + \lambda_2\mathcal{L}_{\mathrm{pp}}, \tag{2}$$

where $\mathcal{L}_{\mathrm{adv}}$ is the **adv**ersarial loss for adding realism to the synthetic images and alleviating artifacts, $\mathcal{L}_{\mathrm{ip}}$ is the **i**dentity **p**erception loss for preserving the identity information, and $\mathcal{L}_{\mathrm{pp}}$ is the **p**ose **p**erception loss for preserving pose and texture information.

$\mathcal{L}_{\mathrm{pp}}$ is a pixel-wise ℓ_1 loss, which is introduced to enforce the pose (*i.e.*, yaw angle) consistency for the synthetic profile face images before and after the refinement via DA-GAN:

$$\mathcal{L}_{\mathrm{pp}} = \frac{1}{W\times H}\sum_{i}^{W}\sum_{j}^{H}|x_{i,j} - \tilde{x}_{i,j}|, \tag{3}$$

where i, j traverse all pixels of x and $\tilde{x}$.

Although $\mathcal{L}_{\mathrm{pp}}$ may bring some over-smooth effects to the refined results, it is still an essential part for both pose and texture information preserving and accelerated optimization.

To add realism to the synthetic images to really benefit face recognition performance, we need to narrow the gap between the distributions of synthetic and real images. An ideal generator will make it impossible to classify a given image as real or refined with high confidence. Meanwhile, preserving the identity information is the essential and critical part for recognition. An ideal generator will generate the refined face images that have small intra-class distance and large inter-class distance in the feature space spanned by the deep neural networks for unconstrained face recognition. These motivate us to use an adversarial pixel-wise discriminator with dual agents.

3.3 Dual-agent discriminator

To incorporate the prior knowledge from the profile faces' distribution and domain knowledge of identities' distribution, we herein introduce a discriminator with dual agents for distinguishing real *v.s.* fake and identities simultaneously. To facilitate this process, we leverage an auto-encoder as the discriminator $D_\phi : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{H\times W\times C}$, which is very simple to avoid typical GAN tricks. It first projects the input real/fake face image into a high-dimensional feature space through several **Conv**olution (Conv) and **F**ully **C**onnected (FC) layers of the encoder, which is then transformed back to the image-level representation through several **Deconv**olution (Deconv) and Conv layers of the decoder, as shown in Fig. 2. ϕ denotes the network parameters for the discriminator. More details are provided in Sec. 4.

One agent of D_ϕ is trained with $\mathcal{L}_{\mathrm{adv}}$ to minimize the Wasserstein distance with a boundary equilibrium regularization term for maintaining a balance between the generator and discriminator losses as first introduced in [31]:

$$\mathcal{L}_{\mathrm{adv}} = \sum_j |y_j - D_\phi(y_j)| - k_t\sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)|, \tag{4}$$

where y denotes the real face image, k_t is a boundary equilibrium regularization term using Proportional Control Theory to maintain the equilibrium $\mathbb{E}[\sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)|] = \gamma\mathbb{E}[\sum_j |y_j - D_\phi(y_j)|]$, and γ is the diversity ratio.

Here k_t is updated by

$$k_{t+1} = k_t + \alpha(\gamma\sum_j |y_j - D_\phi(y_j)| - \sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)|), \tag{5}$$

where α is the learning rate (proportional gain) for k. In essence, Eq. (5) can be viewed as a form of close-loop feedback control in which k_t is adjusted at each step.

$\mathcal{L}_{\mathrm{adv}}$ serves as a supervision to push the refined face image to reside in the manifold of real images. It can avoid the blurry effect, alleviate artifacts and produce visually pleasing results.

The other agent of D_ϕ is trained with $\mathcal{L}_{\mathrm{ip}}$ to preserve the identity discriminability of the refined face images. Specially, we define $\mathcal{L}_{\mathrm{ip}}$ with the proposed enforced cross-entropy loss based on the output from the bottleneck layer of D_ϕ. The enforced cross-entropy loss reduces the intra-class distance while increasing the inter-class distance. Moreover, it helps improve the robustness of the learned representations and address the potential over-fitting issue. We define every column vector of the weights of the bottleneck layer of D_ϕ as an anchor vector a which represents the *center* of each identity in the feature space. Thus, the decision boundary can be derived when the feature vector has the same distance (cosine metric) to several anchor vectors (cluster centers), *i.e.*, $a_i^\top f = a_j^\top f$, where f denotes the feature encoded by the encoder of D_ϕ given the input real/synthetic face image.

However, in such cases, the samples close to the decision boundary would be wrongly classified with a high confidence. A simple yet effective solution is to reduce the intra-class distance while increasing the inter-class distance of the feature vectors, through which the hard samples will be adjusted and re-allocated in the correct decision area. To achieve this goal, we propose to impose a selective attenuation factor as a regularization term to the confidence scores (predictions) of the genuine samples:

$$p_i = \frac{\exp[\tau_t\cdot(a_i^\top f)]}{\sum_j \exp[\tau_t\cdot(a_j^\top f)]}, \tag{6}$$

where p_i denotes the predicted confidence score *w.r.t.* the i^{th} identity, τ_t denotes the selective attenuation factor, and a and f are ℓ_2 normalized to achieve boundary equilibrium during network training. In particular, τ_t in Eqn. (6) is updated by

$\tau_{t+1} = \tau_t \left(1 - \frac{n}{B}\right)^{\alpha}$, where n denotes the batch index, B denotes the total batch number, and α denotes the diversity ratio.

Selective attenuation on the confidence scores of genuine samples in turn increases the corresponding classification losses, which narrows the decision boundary and controls the intra-class affinity and inter-class distance. The predictions of Eqn. (6) are used to compute the multi-class cross-entropy objective function for updating network parameters, which is an enforced optimization scheme:

$$\begin{aligned}\mathcal{L}_{\text{ip}} = &\frac{1}{N}\sum_{j} -Y_j log(p) - (1-Y_j) log(1-p) \\ &+ \frac{1}{N}\sum_{i} -Y_i log(\tilde{p}) - (1-Y_i) log(1-\tilde{p}),\end{aligned} \tag{7}$$

where Y denotes the identity ground truth.

Thus, minimizing $\mathcal{L}_{\text{ip}}$ would encourage deep features of the refined face images belonging to the same identity to be close to each other. If one visualizes the learned deep features in the high-dimensional space, the learned deep features of the refined face image set form several compact clusters and each cluster may be far away from others. Each cluster has a small variance. In this way, the refined face images are enforced with well preserved identity information. We also conduct experiments for illustration.

Using $\mathcal{L}_{\text{ip}}$ alone makes the results prone to annoying artifacts, because searching for a local minimum of $\mathcal{L}_{\text{ip}}$ may go through a path that resides outside the manifold of natural face images. Thus, we combine $\mathcal{L}_{\text{ip}}$ with $\mathcal{L}_{\text{adv}}$ as the final objective function for D_ϕ to ensure that the search resides in that manifold and produces photorealistic and identity-preserving face image:

$$\mathcal{L}_{\text{D}_\phi} = \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{ip}}. \tag{8}$$

3.4 Loss functions for training

The goal of DA-GAN is to use a set of unlabeled real face images y to learn a generator G_θ that adaptively refines a simulated profile face image x. The overall objective function for DA-GAN is

$$\begin{cases}\mathcal{L}_{\text{D}_\phi} = \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{ip}}, \\ \mathcal{L}_{\text{G}_\theta} = (-\mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{ip}}) + \lambda_2 \mathcal{L}_{\text{pp}}.\end{cases} \tag{9}$$

We optimize DA-GAN by alternatively optimizing D_ϕ and G_θ for each training iteration. Similar as in [31], we measure the convergence of DA-GAN by using the boundary equilibrium concept: we can frame the convergence process as finding the closest reconstruction $\sum_j |y_j - D_\phi(y_j)|$ with the lowest absolute value of the instantaneous process error for the Proportion Control Theory $|\gamma \sum_j |y_j - D_\phi(y_j)| - \sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)||$. This measurement can be formulated as

$$\mathcal{L}_{\text{con}} = \sum_j |y_j - D_\phi(y_j)| + |\gamma \sum_j |y_j - D_\phi(y_j)| - \sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)||. \tag{10}$$

$\mathcal{L}_{\text{con}}$ can be used to determine when the network reaches its final state or if the model has collapsed.

4 Experiments

4.1 Experimental settings

4.1.1 Benchmark datasets

Except for synthesizing natural-looking profile view face images, the proposed DA-GAN also aims to generate identity-preserving face images for accurate face-centric analysis with state-of-the-art deep learning models. Thus we evaluate the possibility of

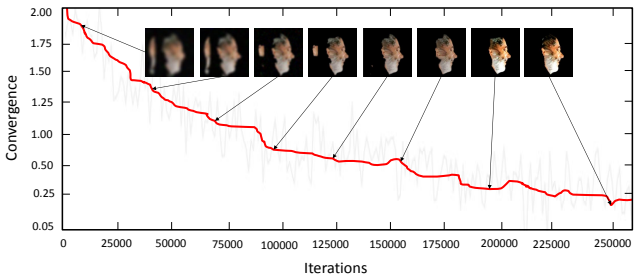


Fig. 3: Quality of refined results *w.r.t.* the network convergence measurement $\mathcal{L}_{\text{con}}$.

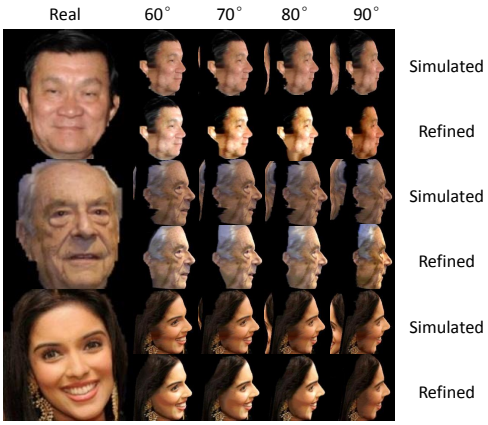


Fig. 4: Refined results of DA-GAN.

"recognition via generation" of DA-GAN on the challenging unconstrained face recognition benchmark datasets IJB-A [7] and CFP [13].

IJB-A [7] contains both images and video frames from 500 subjects with 5,397 images and 2,042 videos that are split into 20,412 frames, namely 11.4 images and 4.2 videos per subject, captured from in-the-wild environment to avoid the near frontal bias and along with protocols for evaluation of both *verification* (1:1 comparison) and *identification* (1:N search) tasks. For training and testing, 10 random splits are provided by each protocol. IJB-A [7] defines the minimal facial representation unit to be a "template" enrolled with multiple face images and/or video frames under extreme conditions of pose, expression, occlusion and illumination. Such problem setting is aligned better with real-world scenario where each subject's appearance is more likely to be captured more than once using different approaches, turning the traditional face recognition problem into a more challenging set-to-set matching problem under extreme conditions in the wild. The verification task requires the evaluation system to determine whether two input face templates are of the same subject or not. At a given threshold, the **R**eceiver **O**perating **C**haracteristic (ROC) analysis measures the **T**rue **A**ccept **R**ate (TAR), which is the fraction of genuine comparisons that correctly exceed the threshold, and the **F**alse **A**ccept **R**ate (FAR), which is the fraction of impostor comparisons that incorrectly exceed the threshold. For identification, the evaluation system needs to determine if the subject matching a probe identity is from a closed set or an open set. For a closed set, the **C**umulative **M**atch **C**haracteristic (CMC) analysis measures the percentage of probe searches returning probe gallery mates within a given Rank. For an open set, at a given threshold, the evaluation system measures the **F**alse **P**ositive **I**dentification **R**ate (FPIR), which is the fraction of comparisons between probe templates and non-mate gallery templates which

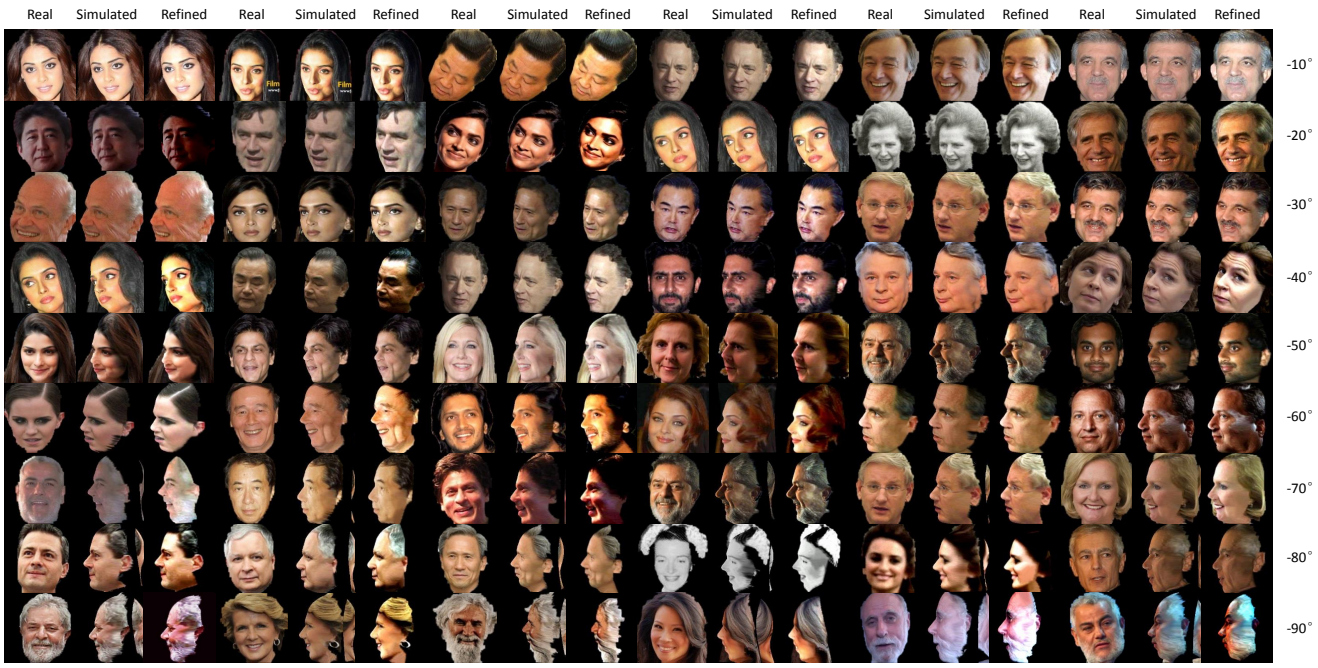


Fig. 5: Refined results of DA-GAN under various poses with yaw angles ranging from -90° to -10° at a stride of 10°.

Fig. 6: Refined results of DA-GAN under various poses with yaw angles ranging from $+10^\circ$ to $+90^\circ$ at a stride of 10°.

corresponds to a match score exceeding the threshold, and the **F**alse **N**egative **I**dentification **R**ate (FNIR), which is the fraction of probe searches that fail to match a mated gallery template above a score of the threshold. More details on the evaluation metrics can be found in [7].

The CFP [13] dataset aims to evaluate the strength of face verification approaches across poses, more specifically, between yaw angle$<10^\circ$ and yaw angle$>60^\circ$. CFP [13] contains 7,000 images of 500 subjects, where each subject has 10 frontal and 4 profile face images. The data are randomly organized into 10 splits, each containing an equal number of frontal-frontal and frontal-profile pairs, with 350 genuine and 350 imposter ones, respectively. Evaluation systems report the mean and standard deviation of accuracy, **E**qual **E**rror **R**ate (EER), and **A**rea **U**nder **C**urve (AUC) over the 10 splits for both frontal-frontal and frontal-profile face verification settings.

4.1.2 *Reproducibility*

The proposed method is implementated by extending the Keras framework [38]. All networks are trained on four NVIDIA GeForce GTX TITAN X GPUs with 12GB memory for each.

4.1.3 *Network architectures*

- Simulator: RAR [9] + iFAN [35] framework (face RoI extraction & saliency prediction & 68 facial landmark detection), 3D MM [10] (profile face image simulation with pre-defined yaw angles).

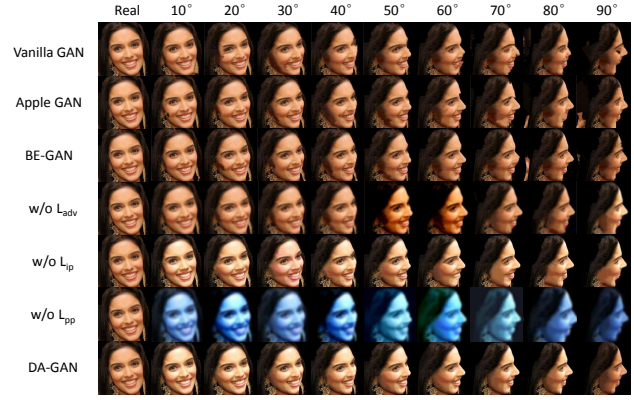


Fig. 7: Qualitative result comparison of DA-GAN with state-of-the-art GANs and three different network settings.

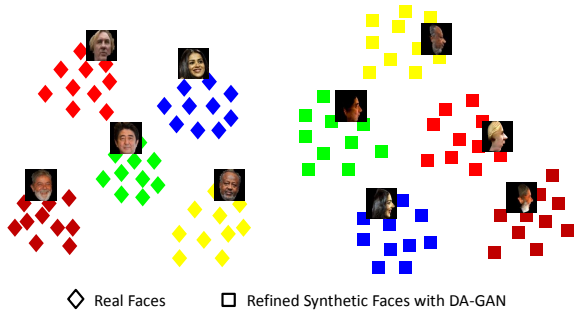


Fig. 8: Feature space of real faces and DA-GAN synthetic faces.

- Generator: Input $224 \times 224 \times 3$, Conv $64 \times 7 \times 7$, ReLU[2], BN[3], $10\times$Residual block (Conv $64 \times 7 \times 7$, ReLU, BN, Conv $64 \times 7 \times 7$, Ele-Sum[4], ReLU, BN), Conv $3 \times 1 \times 1$.
- Discriminator: Input $224 \times 224 \times 3$, Conv $3 \times 3 \times 3$, ReLU, Transition down (Conv $128 \times 3 \times 3$, ReLU, Conv $128 \times 3 \times 3/2$, ReLU, Conv $256 \times 3 \times 3$, ReLU, Conv $256\times3\times3/2$, ReLU, Conv $384\times3\times3$, ReLU, Conv $384\times 3 \times 3/2$, ReLU), Flatten, FC 784, Reshape, Transition up (Conv $128 \times 3 \times 3$, ReLU, Deconv $128 \times 3 \times 3/2$, ReLU, Conv $128 \times 3 \times 3$, ReLU, Deconv $128 \times 3 \times 3/2$, ReLU, Conv $128 \times 3 \times 3$, ReLU, Deconv $128 \times 3 \times 3/2$, ReLU), Conv $3 \times 1 \times 1$, ReLU.
- Deep recognition models: Input $224 \times 224 \times 3$, SE-ResNeXt-152 (cardinality $= 32$) [15], [16] & GoogleNet-BN [17] (model fusion), template-adapted **S**upport **V**ector **M**achine (SVM) [1] (metric learning).

The overview of our proposed "recognition via generation" framework is illustrated in Fig. 2. We transfer-learn two state-of-the-art deep neural networks, SE-ResNeXt-152 [15], [16] and GoogleNet-BN [17], from source domain (MS-Celeb-1M [41] and VGGFace2 [42], with noise and overlapping parts removed with IJB-A [7]/CFP [13]) to target domain of IJB-A [7]/CFP [13] extended by DA-GAN. We combine the complementary two-view information (learned deep features) from the SE-ResNeXt-152 [15], [16] and GoogleNet-BN [17] models to train template-adapted SVMs [1]. The resulted margins are robust and discriminative for unconstrained face recognition.

2. ReLU is short for "**Re**ctified **L**inear **U**nits" [39].
3. BN is short for "**B**atch **N**ormalization" [40].
4. Ele-Sum is short for "**ele**ment-wise **sum**mation".

4.1.4 Training details

- DA-GAN: 1) Extract face RoIs from the available training data, perform saliency prediction (*i.e.*, face/background segmentation), and estimate 68 facial landmark points using the RAR [9] + iFAN [35] framework. 2) Simulate profile faces with pre-defined yaw angles $\in \{\pm10, \pm20, \pm30, \pm40, \pm50, \pm60, \pm70, \pm80, \pm90\}$ using 3D MM [10]. 3) Train DA-GAN using Adam with mini-batch (FC 333 with enforced Softmax appended to the output of the bottleneck layer of D_ϕ for $\mathcal{L}_{\text{ip}}$ during training); set the mini-batch size to 16; $W = 224$, $H = 224$, $C = 3$; initialize DA-GAN using vanishing residuals; set an initial learning rate to 5×10^{-5}, decaying by a factor of 2 when $\mathcal{L}_{\text{con}}$ stalls; set the weight decay to 5×10^{-4}; set $k_0 = 0$; $\lambda_1 = 2.5 \times 10^{-2}$, $\lambda_2 = 3 \times 10^{-2}$, $\alpha = 1 \times 10^{-3}$, $\gamma = 5 \times 10^{-1}$; alternatively optimize discriminator D_ϕ, generator G_θ and update k_t for each mini-batch.
- Deep recognition models: 1) Set the mini-batch size to 256; $W = 224$, $H = 224$, $C = 3$; set an initial learning rate to 0.01 that is divided by 10 every 30 epoches; set the weight decay to 1×10^{-4}; set the momentum to 0.9. 2) Pre-process the MS-Celeb-1M [41] data, including overlapping part removal with IJB-A [7]/CFP [13] and face RoI extraction, resulting in 4,356,052 face images for 53,317 subjects in total. 3) Train SE-ResNeXt-152 (cardinality $= 32$) [15], [16] & GoogleNet-BN [17] using **S**tochastic **G**radient **D**escent (SGD) on the cleaned MS-Celeb-1M [41] data. 4) Set the mini-batch size to 256; $W = 224$, $H = 224$, $C = 3$; set an initial learning rate to 0.001 that is divided by 10 every 30 epoches; set the weight decay to 1×10^{-4}; set the momentum to 0.9. 5) Pre-process the VGGFace2 [42] data, including overlapping part removal with IJB-A [7]/CFP [13] and face RoI extraction, resulting in 3.31 million face images for 9,131 subjects in total. 6) Fine-tune SE-ResNeXt-152 (cardinality $= 32$) [15], [16] & GoogleNet-BN [17] using **S**tochastic **G**radient **D**escent (SGD) on the cleaned VGGFace2 [42] data. 7) Reset the learning rate to 0.0001 that is divided by 10 every 10 epoches. 8) Inject the refined profile faces into IJB-A [7]/CFP [13] training data and fine-tune the pre-trained deep recognition models.
- Template-adapted SVM models for IJB-A [7]: 1) Concatenate the learned pose-invariant features from the penultimate layers of deep recognition models ($\mathbb{R}^{2048}$ C-Sum[5] $\mathbb{R}^{1024} \mapsto \mathbb{R}^{3072}$). 2) Train template-adapted SVM models similarly as introduced in [1].

Formally, the template-adapted SVMs are learned by optimizing the following ℓ_2-regularized objective function:

$$\mathcal{L}_{\text{SVM}} = \min_w \frac{1}{2} w^T w + \lambda_+ \sum_{i=1}^{N_+} \max\left[0, 1 - y_i w^T f_F\left(\mathbf{x}_i\right)\right]^2 + \lambda_- \sum_{j=1}^{N_-} \max\left[0, 1 - y_j w^T f_F\left(\mathbf{x}_j\right)\right]^2 , \tag{11}$$

where $f_F(\cdot)$ denotes the non-linear function learned by our deep recognition models, x denotes the face media, w denotes the

5. C-Sum is short for "**c**oncatenate".

Fig. 9: Verification result analysis for best matched cases (left) and best non-matched cases (right) on IJB-A [7] split1. Best viewed in the original zoomed-in pdf.

weights including bias term, $y_i \in \{-1, 1\}$ denotes the label showing whether the current sample is negative or positive, N_+ is the number of positive samples, N_- is the number of negative ones, $N_- \gg N_+$, the constraint for negative samples $\lambda_- = C\frac{N_+ + N_-}{2N_-}$, the constraint for positive samples $\lambda_+ = C\frac{N_+ + N_-}{2N_+}$, and C denotes a trade-off factor, which is set to 20 in our method.

Recall a template contains both face images and/or video frames, with large variance in terms of media modality, pose, expression, occlusion and illumination. Recently, several set-based face recognition methods have been proposed [2], [43], [44], [45]. They generally adopt the following two strategies to obtain set-level face representation. One is to learn a set of image-level face representations from each face medium in the set individually [2], [44], and use all the information for following face recognition. Such a strategy is obviously computationally expensive as it needs to perform exhaustive pairwise matching and is fragile to outlier faces captured under unusual conditions. The other strategy is to aggregate face representations across the set through average or max pooling and generate single representation for each set [3], [45], which obviously suffers from information loss. Therefore, neither of the strategies can effectively solve the set-based face recognition problems. In order to better address the underlying distracting factors within each template, we split each template into several sub-templates[6] according to the prior information on the media source (*e.g.*, image/video). In particular, for the deep features from a video sequence, we perform mean encoding to generate the corresponding representation while keeping the deep feature of each image unchanged.

Let t_j^V be the mean encoding of the j^{th} video sequence. Then

$$t_j^V = \frac{1}{N_j^V}\sum_{i=1}^{N_j^V} f_F\left(\mathbf{x}_i\right), \tag{12}$$

where N_j^V denotes the number of frames in the j^{th} video sequence and $\mathbf{x}_i$ denotes the i^{th} frame of video j.

Thus, the representations for the a^{th} template can be expressed as

$$T_a = \left\{t_i^I, ..., t_{N_a}^V\right\}, \tag{13}$$

where t_i^I denotes the sub-template for the i^{th} image and $t_{N_a}^V$ denotes the sub-template for the N_a^{th} video.

The media-level deep features are further ℓ_2-normalized for training template-adapted SVMs [1]. For verification, the positive sample of template specific SVM is a probe template, and the

6. We observe similar recognition performance between our "sub-template" matching strategy and the "exhaustive" matching strategy in evaluations on IJB-A [7] (within $\pm 0.4\%$ variation). We hence choose the "sub-template" matching strategy for a good trade-off between high accuracy and efficiency.

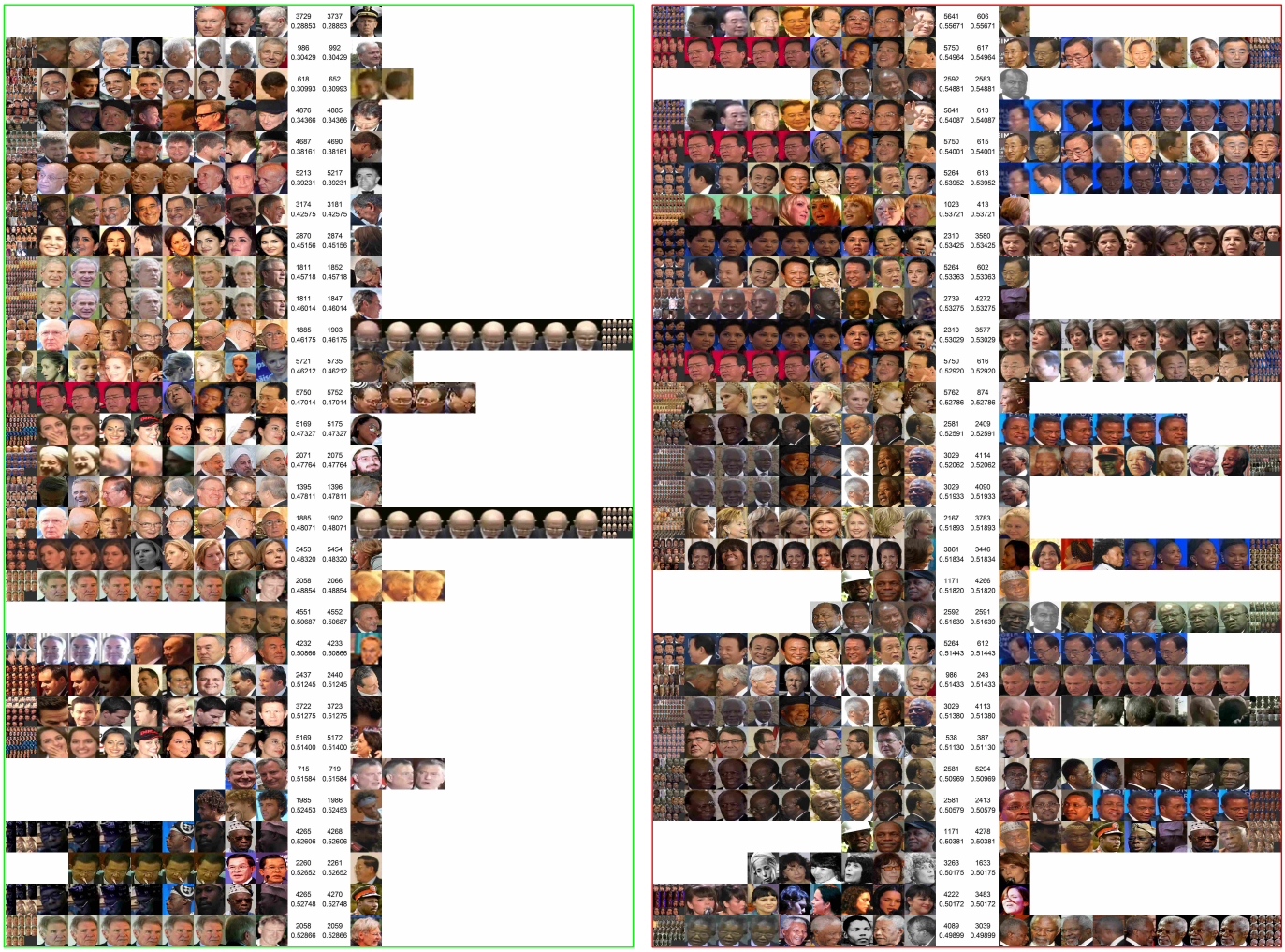


Fig. 10: Verification result analysis for worst matched cases (left) and worst non-matched cases (right) on IJB-A [7] split1. Best viewed in the original zoomed-in pdf.

large-scale negative samples consist of the whole training set. For identification, the probe template specific SVMs adopt the whole training set as the large-scale negative samples; for gallery template specific SVM, other gallery templates and the whole training set are bundled together as the large-scale negative samples.

Based on **O**ne-**S**hot **S**imilarity (OSS), we compute the fine-grained similarity between two sub-template representations p and q via $s\left(p,q\right)=\frac{1}{2}[\mathcal{P}\left(q\right)+\mathcal{Q}\left(p\right)]$, where $\mathcal{P}(\cdot)$ denotes the trained probe template specific SVM model and $\mathcal{Q}(\cdot)$ indicates the trained gallery template specific SVM model.

As described in Eq. (13), a template may contain various numbers of sub-templates. Thus, finally we merge the resulting multiple matching scores into a single measurement to determine the face identity for each template pair:

$$s\left(T_a,T_b\right)=\frac{\sum\limits_{t_i\in T_a,t_j\in T_b}s\left(t_i,t_j\right)e^{\beta\, s\left(t_i,t_j\right)}}{\sum\limits_{t_i\in T_a,t_j\in T_b}e^{\beta\, s\left(t_i,t_j\right)}},\tag{14}$$

where β is a bandwidth factor, and we set it to 0 in our method.

4.2 Results and discussions

4.2.1 *Qualitative results on IJB-A – DA-GAN*

In order to illustrate the compelling perceptual results generated by the proposed DA-GAN, we first visualize the quality of refined results *w.r.t.* the network convergence measurement $\mathcal{L}_{\text{con}}$ in Fig. 3. Obviously, our DA-GAN ensures a fast yet stable convergence through the carefully designed optimization scheme and boundary equilibrium regularization term. The network convergence measurement $\mathcal{L}_{\text{con}}$ correlates well with image fidelity.

Most of previous works [8], [9], [10] on profile view synthesis address this problem within a pose range of $\pm 60°$, since it is commonly believed with a pose exceeding $60°$, it is difficult for a model to generate faithful profile view images. Also, our simulator is good at normalizing small pose faces but suffers severe artifacts and texture loss under large poses (*e.g.*, yaw angles $\in \{[-90°,-60°]\cup[+60°,+90°]\}$), as shown in Fig. 4 *row*. 1 for each subject. However, with enough training data and proper architecture and objective function design of the proposed DA-GAN, it is in fact feasible to further refine such synthetic profile face images under very large poses for high-quality natural-looking results generation, as shown in Fig. 4 *row*. 2 for each subject. Compared with the raw simulated faces, the refined results by DA-GAN present good photorealistic quality. We visualize the high-resolution refined results of DA-GAN under various poses, with yaw angles ranging from $\pm 10°$ to $\pm 90°$ at a stride of $10°$, in Fig. 5 and Fig. 6 to verify the remarkable perceptual quality of DA-GAN. As can be seen, DA-GAN is able to adaptively remove artifacts (*e.g.*, face fragments and black holes) from the

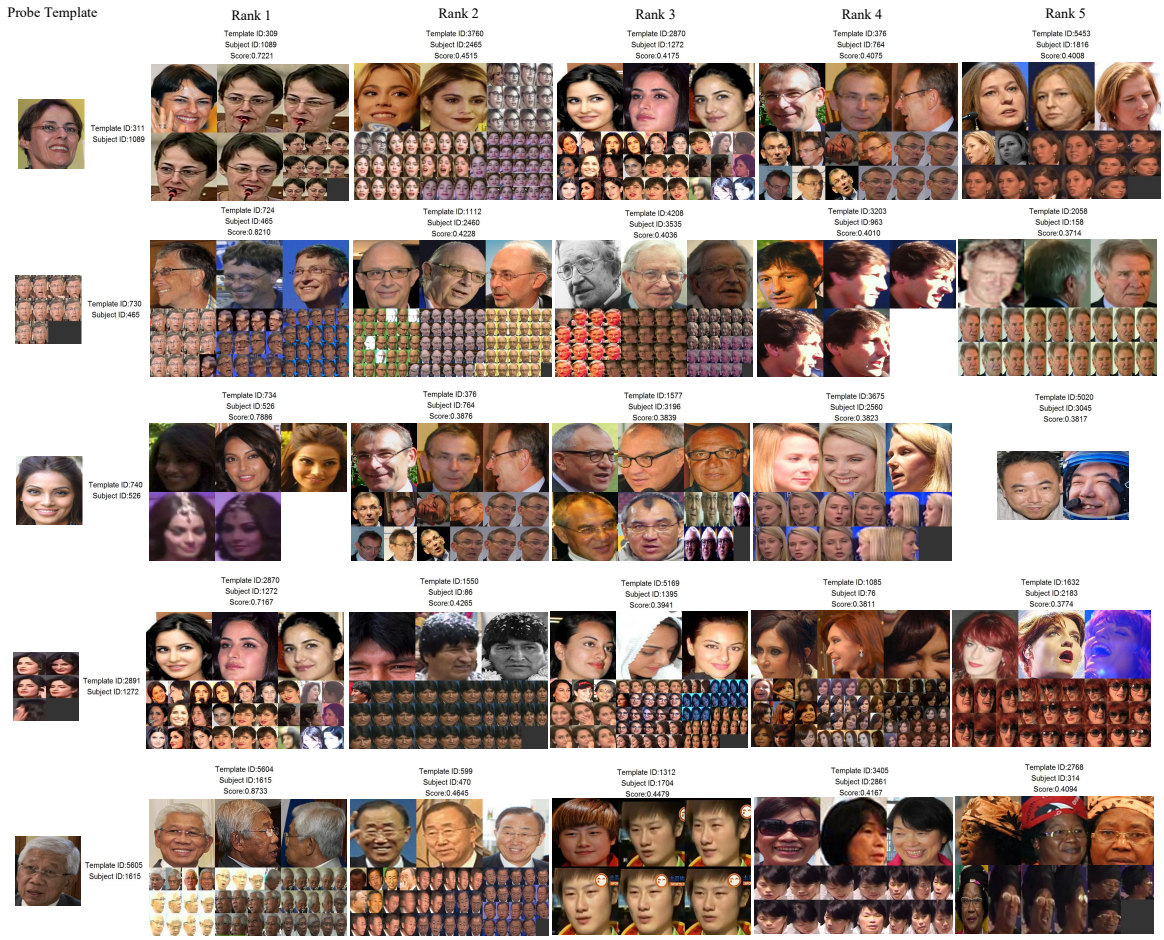


Fig. 11: Identification results analysis on IJB-A [7] split1. Best viewed in the original zoomed-in pdf.

simulator, stitch fragments and compensate texture loss in terms of facial details and color realism, especially for large poses. The refined faces of DA-GAN present more intuitively photorealistic and natural characteristics.

To verify the superiority of DA-GAN and the contribution of each component, we also compare the qualitative results produced by the vanilla GAN [12], Apple GAN [34], BE-GAN [31], and three variations of DA-GAN in terms of w/o $\mathcal{L}_{\rm adv}$, $\mathcal{L}_{\rm ip}$, $\mathcal{L}_{\rm pp}$ in each case. As shown in Fig. 7, inference without $\mathcal{L}_{\rm ip}$ deviates from the true appearance seriously, and the synthesis without $\mathcal{L}_{\rm adv}$ tends to be very blurry, while the results without the $\mathcal{L}_{\rm pp}$ sometimes show blurry and unnatural effects with strange artifacts/color involved. Compared with vanilla GAN [12], Apple GAN [34] and BE-GAN [31], which all fail for poses larger than 60°, our DA-GAN achieves good identity-preserving quality while producing photorealistic synthesis.

To evaluate the effectiveness for identity-preserving of our DA-GAN, we further use t-SNE [57] to visualize the deep features of both refined profile faces and real faces in a 2D space in Fig. 8. The refined profile face images present small intra-class distance and large inter-class distance, which is similar to those of real faces. This reveals that DA-GAN ensures well preserved identity information with the auxiliary agent for $\mathcal{L}_{\rm ip}$.

4.2.2 *Quantitative results on IJB-A – "recognition via generation"*

To quantitatively verify the superiority of "recognition via generation" of DA-GAN, we conduct unconstrained face recognition (*i.e.*, verification and identification) on IJB-A [7] dataset with five different settings. In the five settings, the pre-trained deep recognition models are respectively fine-tuned on the original training data of each split without extra data (baseline 1: b1), the original training data of each split with extra synthetic faces by our simulator (baseline 2: b2), the original training data of each split with extra simulated faces after fragment removal[7] (baseline 3: b3), the original training data of each split with extra refined faces by our preliminary DA-GAN employing standard cross-entropy loss as $\mathcal{L}_{\rm ip}$, and the training data of each split augmented by extra refined faces by DA-GAN using the proposed enforced cross-entropy loss as $\mathcal{L}_{\rm ip}$ (our method: "recognition via generation" framework based on enhanced DA-GAN, DA-GAN$_{2.0}$ for short). The performance comparison of DA-GAN with the four baselines and other state-of-the-arts on IJB-A [7] unconstrained face verification and identification protocols are given in Table 1 and 2.

We can observe that even with extra training data, b2 presents inferior performance to b1 for all metrics of both face verification

7. We remove the fragments from each simulated face image where a *fragment* is considered as a small piece less than 30% that has broken off from the main face region.

TABLE 1: Performance comparison of DA-GAN with state-of-the-arts on IJB-A [7] verification protocol. For all metrics, a higher number means better performance. The results are averaged over 10 testing splits. Symbol "-" implies that the result is not reported for that method. Standard deviation is not available for some methods. The results offered by our proposed method are highlighted in bold.

Method	TAR@FAR=0.10	TAR@FAR=0.01	TAR@FAR=0.001	TAR@FAR=0.0001
OpenBR [7]	0.433±0.006	0.236±0.009	0.104±0.014	-
GOTS [7]	0.627±0.012	0.406±0.014	0.198±0.008	-
Pooling faces [43]	0.631	0.309	-	-
LSFS [46]	0.895±0.013	0.733±0.034	0.514±0.060	-
Deep Multi-pose [47]	0.911	0.787	-	-
$\text{DCNN}_{\text{manual}}$+metric [48]	0.947±0.011	0.787±0.043	-	-
Triplet Similarity [3]	0.945±0.002	0.790±0.030	0.590±0.050	-
VGG-Face [49]	-	0.805±0.030	-	-
PAMs [2]	-	0.826±0.018	0.652±0.037	-
$\text{DCNN}_{\text{fusion}}$ [45]	0.967±0.009	0.838±0.042	-	-
FF-GAN [26]	-	0.852±0.010	0.663±0.033	-
DR-GAN [25]	-	0.831±0.017	0.699±0.029	-
Masi *et al.* [50]	-	0.886	0.725	-
Chen *et al.* [51]	0.968±0.005	0.889±0.016	0.760±0.038	-
Triplet Embedding [3]	0.964±0.005	0.900±0.010	0.813±0.002	-
All-In-One [52]	0.976±0.004	0.922±0.010	0.823±0.020	-
Template Adaptation [1]	0.979±0.004	0.939±0.013	0.836±0.027	-
NAN [4]	0.979±0.004	0.941±0.008	0.881±0.011	-
ℓ_2-softmax [53]	0.984±0.002	0.970±0.004	0.943±0.005	0.909±0.007
b1	0.989±0.003	0.963±0.007	0.920±0.006	-
b2	0.978±0.003	0.950±0.009	0.901±0.008	-
b3	0.980±0.003	0.956±0.009	0.907±0.008	-
DA-GAN	0.991±0.003	0.976±0.007	0.930±0.005	-
$\text{DA-GAN}_{2.0}$	**0.995±0.001**	**0.989±0.002**	**0.973±0.005**	**0.946±0.010**

and identification. Even though b3 slightly improves b2 due to additional post-processing, it still has a clear margin compared with b1. This demonstrates that naively learning from synthetic images can be problematic due to a gap between synthetic and real image distributions – synthetic data are often not realistic enough with artifacts and severe texture loss, misleading the network to overfit to fake information only contained in synthetic images and fail to generalize well on real data. In contrast, with the injection of photorealistic and identity-preserving faces generated by DA-GAN without extra human annotation efforts, our preliminary method outperforms b1 by 1.00% for TAR@FAR=0.001 of verification and 1.50% for FNIR@FPIR=0.01, 0.50% for Rank1 of identification; with the introduction of the enforced cross-entropy optimization strategy, our $\text{DA-GAN}_{2.0}$ further achieves improvements of 5.30% for TAR @ FAR=0.001 of verification and 6.40% for FNIR@FPIR=0.01, 2.40% for Rank1 of identification. Moreover, $\text{DA-GAN}_{2.0}$ outperforms the 2^{nd}-best method by 3.70% for TAR@FAR=0.0001 of verification and 2.40% for FNIR@FPIR=0.01, 1.70% for Rank1 of identification. DA-GAN is the foundation of our winning entry to the NIST IJB-A face recognition competition in which we secured the 1^{st} places on the tracks of verification and identification[8]. This well verified the potential of synthetic face images by our DA-GAN on the large-scale and challenging unconstrained face recognition problem.

Finally, we visualize the verification and identification closed set results for IJB-A [7] split1 to gain insights into unconstrained face recognition with the proposed "recognition via generation" framework based on DA-GAN.

For face verification, after computing the similarities for all pairs of probe and reference sets, we sort the results into a ranking list. Each row shows a probe and reference template pair. The original templates within IJB-A [7] contain from one to dozens of media. Up to eight individual media are shown, with the last space showing a mosaic of the remaining media in the template. Between the templates are the template IDs for probe and reference as well as the best matched and best non-matched similarities. Fig. 9 left shows the best matched cases. In the top-30 scoring correct matches, we note that every reference template contains dozens of media. The probe templates either contain one or dozens of well-matched media. Fig. 9 right illustrating the best non-matched cases shows the most certain non-mates, again often involving large templates with enough guidance from the relevant information of the same subject. Fig. 10 left shows the worst matched cases, representing failed matching. The thirty lowest matched results from single-medium probe sets are all under extremely challenging unconstrained conditions, which cannot be solved even using the specific operations designed in our "recognition via generation" framework. Fig. 10 right illustrating the worst non-matched cases highlights the understandable error, representing impostors in challenging modalities.

For face identification, Fig. 11 *col.* 1 shows the query images from probe templates. Fig. 11 *col.* 2-6 show the corresponding top-5 queried gallery templates. For each template, we provide template ID, subject ID and similarity score. As can be seen, our approach always gives successful searching in Rank1, which well proves its effectiveness for generic transfer learning and face-centric analysis. It would be interesting to apply DA-GAN to other transfer learning applications in future.

4.2.3 *Quantitative results on CFP – "recognition via generation"*

To further verify the effectiveness and generalizability of our DA-GAN for unconstrained face recognition in the wild, we compare its recognition performance with human performance and other state-of-the-arts on the CFP [13] benchmark dataset in Table 3. DA-GAN outperforms human performance and other state-of-the-arts under both frontal-profile setting and frontal-frontal setting. In particular, for frontal-frontal cases, DA-GAN

8. We submitted our results for both verification and identification protocols to NIST IJB-A 2017 face recognition competition committee on 29th, March, 2017. We received the official notification on our top performance on both tracks on 26th, Apirl, 2017. The IJB-A benchmark dataset, relevant information and leaderboard can be found at https://www.nist.gov/programs-projects/face-challenges.

TABLE 2: Performance comparison of DA-GAN with state-of-the-arts on IJB-A [7] identification protocol. For FNIR metric, a lower number means better performance. For the other metrics, a higher number means better performance. The results are averaged over 10 testing splits. Symbol "-" implies that the result is not reported for that method. Standard deviation is not available for some methods. The results offered by our proposed method are highlighted in bold.

Method	FNIR@FPIR=0.10	FNIR@FPIR=0.01	Rank1	Rank5	Rank10
OpenBR [7]	0.851±0.028	0.934±0.017	0.246±0.011	0.375±0.008	-
GOTS [7]	0.765±0.033	0.953±0.024	0.433±0.021	0.595±0.020	-
B-CNN [44]	0.659±0.032	0.857±0.027	0.588±0.020	0.796±0.017	-
Pooling faces [43]	-	-	0.846	0.933	0.951
LSFS [46]	0.387±0.032	0.617±0.063	0.820±0.024	0.929±0.013	-
Deep Multi-pose [47]	0.250	0.480	0.846	0.927	0.947
$\text{DCNN}_{\text{manual}}$+metric [48]	-	-	0.852±0.018	0.937±0.010	0.954±0.007
Triplet Similarity [3]	0.246±0.014	0.444±0.065	0.880±0.015	0.950±0.007	0.974±0.006
VGG-Face [49]	0.330±0.031	0.539±0.077	0.913±0.011	-	0.981±0.005
PAMs [2]	-	-	0.840±0.012	0.925±0.008	0.946±0.007
DR-GAN [25]	-	-	0.901±0.014	0.953±0.011	-
FF-GAN [26]	-	-	0.902±0.006	0.954±0.005	-
$\text{DCNN}_{\text{fusion}}$ [45]	0.210±0.033	0.423±0.094	0.903±0.012	0.965±0.008	0.977±0.007
Masi *et al.* [50]	-	-	0.906	0.962	0.977
Triplet Embedding [3]	0.137±0.014	0.247±0.030	0.932±0.010	-	0.977±0.005
Template Adaptation [1]	0.118±0.016	0.226±0.049	0.928±0.001	0.977±0.004	0.986±0.003
Chen *et al.* [51]	0.164±0.010	0.346±0.001	0.942±0.008	0.980±0.005	0.988±0.003
All-In-One [52]	0.113±0.014	0.208±0.020	0.947±0.008	-	0.988±0.003
NAN [4]	0.083±0.009	0.183±0.041	0.958±0.005	0.980±0.005	0.986±0.003
Hayat *et al.* [54]	0.040±0.010	0.114±0.041	0.964±0.008	-	1.000±0.000
ℓ_2-softmax [53]	0.044±0.006	0.085±0.041	0.973±0.005	-	0.988±0.003
b1	0.068±0.010	0.125±0.035	0.966±0.006	0.987±0.003	-
b2	0.108±0.008	0.179±0.042	0.960±0.007	0.982±0.004	-
b3	0.101±0.008	0.170±0.042	0.963±0.007	0.984±0.004	-
DA-GAN	0.051±0.009	0.110±0.039	0.971±0.007	0.989±0.003	-
$\text{DA-GAN}_{2.0}$	**0.018±0.003**	**0.061±0.040**	**0.990±0.002**	**0.995±0.003**	**0.997±0.003**

TABLE 3: Performance comparison of DA-GAN with state-of-the-arts on CFP [13]. For EER metric, a lower number means better performance. For the other metrics, a higher number means better performance. The results are averaged over 10 testing splits. The results offered by our proposed method are highlighted in bold.

Method	Frontal-Profile			Frontal-Frontal		
	Acc	EER	AUC	Acc	EER	AUC
FV+DML [13]	58.47±3.51	38.54±1.59	65.74±2.02	91.18±1.34	8.62±1.19	97.25±0.60
LBP+Sub-SML [13]	70.02±2.14	29.60±2.11	77.98±1.86	83.54±2.40	16.00±1.74	91.70±1.55
HoG+Sub-SML [13]	77.31±1.61	22.20±1.18	85.97±1.03	88.34±1.33	11.45±1.35	94.83±0.80
FV+Sub-SML [13]	80.63±2.12	19.28±1.60	88.53±1.58	91.30±0.85	8.85±0.74	96.87±0.39
Deep Features [13]	84.91±1.82	14.97±1.98	93.00±1.55	96.40±0.69	3.48±0.67	99.43±0.31
Triplet Embedding [3]	89.17±2.35	8.85±0.99	97.00±0.53	96.93±0.61	2.51±0.81	99.68±0.16
Chen *et al.* [55]	91.97±1.70	8.00±1.68	97.70±0.82	98.41±0.45	1.54±0.43	99.89±0.06
DR-GAN [25]	93.41±1.17	6.45±0.16	97.96±0.06	97.84±0.79	2.22±0.09	99.72±0.02
P-CNN [56]	94.39±1.17	5.94±0.11	98.36±0.05	97.79±0.40	2.48±0.07	99.71±0.02
Human	94.57±1.10	5.02±1.07	98.92±0.46	96.24±0.67	5.34±1.79	98.19±1.13
DA-GAN	**95.96±0.85**	**4.61±1.03**	**99.00±0.40**	**99.48±0.36**	**0.62±0.39**	**99.97±0.06**

reduces the EER of the 2^{nd}-best by around 0.92%. For more challenging frontal-profile cases, DA-GAN consistently outperforms the human performance and other state-of-the-arts. In particular, DA-GAN reduces the EER by 0.41% compared with human performance and improves the accuracy by 1.33% over the 2^{nd}-best. This shows that the synthesized profile faces by DA-GAN are photorealistic with well-preserved identity information. Such synthetic data can be utilized to augment limited real training data to balance pose distribution, from which discriminative and robust face recognition models can be learned to achieve top performance even for extreme unconstrained conditions.

5 Conclusion

We proposed a novel **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) for photorealistic and identity-preserving profile face synthesis. DA-GAN combines prior knowledge from data distribution (adversarial training) and domain knowledge of faces (pose and identity perception loss) to exactly recover the information lost inherently in projecting a 3D face into the 2D image space. DA-GAN can be optimized in a fast yet stable way with an imposed boundary equilibrium regularization term that balances the power of the discriminator against the generator. One promising potential application of the proposed DA-GAN is for solving generic transfer learning problems more effectively. Qualitative and quantitative experiments verify the possibility of our "recognition via generation" framework, which achieved the top performance on the large-scale and challenging NIST IJB-A and CFP unconstrained face recognition benchmarks without extra human annotation efforts. Based on DA-GAN, we won the 1^{st} places on verification and identification tracks in NIST IJB-A face recognition competition. It would be interesting to apply DA-GAN for other transfer learning applications in future.

Acknowledgments

Jian Zhao's research was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

Junliang Xing's research was partially supported by the National Science Foundation of Chian 61672519.

Jiashi Feng's research was partially supported by National University of Singapore startup R-263-000-C08-133, MOE Tier-

I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

The authors would like to thank Yu Cheng (Nanyang Technological University), Yi Cheng, Yan Xu, Jayashree Karlekar, Sugiri Pranata and Shengmei Shen (Core Technology Group, Learning & Vision, Panasonic R&D Center Singapore) for helpful discussions.

REFERENCES

[1] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *FG*, 2017, pp. 1–8.
[2] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *CVPR*, 2016, pp. 4838–4846.
[3] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *BTAS*, 2016, pp. 1–8.
[4] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *CVPR*, 2017, pp. 4362–4371.
[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.
[6] J. Li, J. Zhao, F. Zhao, H. Liu, J. Li, S. Shen, J. Feng, and T. Sim, "Robust face recognition with deep multi-view representation learning," in *ACM MM*, 2016, pp. 1068–1072.
[7] 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 *CVPR*, 2015, pp. 1931–1939.
[8] S. Xiao, L. Liu, X. Nie, J. Feng, A. A. Kassim, and S. Yan, "A live face swapper," in *ACM MM*, 2016, pp. 691–692.
[9] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *ECCV*, 2016, pp. 57–72.
[10] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li, "Discriminative 3d morphable model fitting," in *FG*, vol. 1, 2015, pp. 1–8.
[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.
[13] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *WACV*, 2016, pp. 1–9.
[14] 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 *NIPS*, 2017, pp. 65–75.
[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 5987–5995.
[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.
[18] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015, pp. 4295–4304.
[19] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *CVPR*, 2015, pp. 787–796.
[20] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikainen, "A survey of recent advances in texture representation," *arXiv preprint arXiv:1801.10324*, 2018.
[21] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: taxonomy and experimental study," *PR*, vol. 62, pp. 135–160, 2017.
[22] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *ICCV*, 2015, pp. 3871–3879.
[23] L. Liu, P. Fieguth, G. Zhao, and M. Pietikäinen, "Extended local binary pattern fusion for face recognition," in *ICIP*, 2014, pp. 718–722.
[24] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *ICCV*, 2017, pp. 2439–2448.
[25] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, vol. 3, no. 6, 2017, p. 7.
[26] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *ICCV*, 2017, pp. 3990–3999.
[27] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *CVPR*, 2015, pp. 676–684.
[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[29] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.
[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[31] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.
[32] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016, pp. 2172–2180.
[33] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv:1610.09585*, 2016.
[34] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training." in *CVPR*, vol. 2, no. 4, 2017, p. 5.
[35] J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim, "Integrated face analytics networks through cross-dataset hybrid training," in *ACM MM*, 2017, pp. 1531–1539.
[36] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *CVPR*, 2017, pp. 6757–6765.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[38] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.
[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[41] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016, pp. 87–102.
[42] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *FG*, 2018, pp. 67–74.
[43] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: template based face recognition with pooled face images," in *CVPRW*, 2016, pp. 59–67.
[44] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear cnns," in *WACV*, 2016, pp. 1–9.
[45] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *WACV*, 2016, pp. 1–9.
[46] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," in *ICB*, vol. 4, 2015.
[47] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan *et al.*, "Face recognition using deep multi-pose representations," in *WACV*, 2016, pp. 1–9.
[48] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *CVPRW*, 2015, pp. 118–126.
[49] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
[50] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *ECCV*, 2016, pp. 579–596.
[51] J.-C. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C.-H. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Unconstrained still/video-based face verification with deep convolutional neural networks," *IJCV*, vol. 126, no. 2-4, pp. 272–291, 2018.
[52] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *FG*, 2017, pp. 17–24.
[53] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.
[54] M. Hayat, S. H. Khan, N. Werghi, and R. Goecke, "Joint registration and representation learning for unconstrained face identification," in *CVPR*, 2017, pp. 2767–2776.

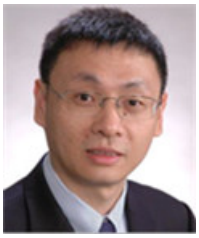
[55] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, "Fisher vector encoded deep convolutional features for unconstrained face verification," in *ICIP*, 2016, pp. 2981–2985.
[56] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *T-IP*, 2017.
[57] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Jian Zhao received the Bachelors degree from Beihang University in 2012, and the Masters degree from the National University of Defense Technology in 2014. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, National University of Singapore. His research is focused on developing deep neural network models and algorithms for human-centric image understanding, applied to face recognition, image generation and human parsing. He has published several cutting-edge projects on unconstrained/large-scale/low-shot face verification/identification and human parsing. He has won the top-3 awards several times on world-wide competitions on face recognition, human parsing and pose estimation.

Lin Xiong received the Bachelors degree from Shanxi University of Science & Technology in 2003, and the Ph.D. degree with School of Electronic Engineering, Xidian University, China, in 2014. He is currently a research engineer of Learning & Vision, Core Technology Group, Panasonic R&D Center Singapore, Singapore. His research interests include unconstrained/large-scale face recognition, person re-identification, deep learning architecture engineering, transfer learning, Riemannian manifold optimization, sparse and low-rank matrix factorization.

Jianshu Li is currently a Ph.D. candidate in School of Computing, National University of Singapore, advised by Prof. Terence Sim and Assoc. Prof. Shuicheng Yan. His research interest is mainly focused on computer vision and image understanding, particularly face and human analytics, semantic segmentation and object detection.

Xing Junliang received his dual B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Dr. Xing was the recipient of Google Ph.D. Fellowship 2011, the Excellent Student Scholarships at Xi'an Jiaotong University from 2004 to 2007 and at Tsinghua University from 2009 to 2011. He has published more than 70 papers on international journals and conferences. His current research interests mainly focus on computer vision problems related to faces and humans.



Shuicheng Yan is currently the Vice-President and the Chief Scientist with Qihoo 360 Technology Company Ltd., and the Head of the 360 Artificial Intelligence Institute. He is also a tenured Associate Professor with the National University of Singapore. He has authored/co-authored over 500 high quality technical papers, with Google Scholar citation over 25 000 times and an h-index 70. His research areas include computer vision, machine learning, and multimedia analysis. He is an IAPR Fellow and the ACM Distinguished Scientist. His team received seven times winner or honorable-mention prizes in five years over PASCAL, VOC, and ILSVRC competitions, which are core competitions in the field of computer vision, along with over ten times the Best (student) Paper Awards and especially a Grand Slam with the ACM MM, the top conference in the field of multimedia, including the Best Paper Award, the Best Student Paper Award, and the Best Demo Award. He is a TR Highly Cited Researcher of 2014, 2015, and 2016.

Jiashi Feng received the Ph.D. degree from the National University of Singapore (NUS) in 2014. He was a Post-Doctoral Research Follow with the University of California, Berkeley. He joined NUS as a Faculty Member, where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering. His research areas include computer vision, machine learning, object recognition, detection, segmentation, robust learning and deep learning.