Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

Jian Zhao[1,2*†] **Lin Xiong**[3] **Karlekar Jayashree**[3] **Jianshu Li**[1] **Fang Zhao**[1]
Zhecan Wang[4†] **Sugiri Pranata**[3] **Shengmei Shen**[3]
Shuicheng Yan[1,5] **Jiashi Feng**[1]
[1]National University of Singapore [2]National University of Defense Technology
[3] Panasonic R&D Center Singapore [4] Franklin. W. Olin College of Engineering
[5] Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu {lin.xiong, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com
zhecan.wang@students.olin.edu {elezhf, eleyans, elefjia}@u.nus.edu

Abstract

Synthesizing realistic profile faces is promising for more efficiently training deep pose-invariant models for large-scale unconstrained face recognition, by populating samples with extreme poses and avoiding tedious annotations. 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 specifically 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 and texture, preserve identity and stabilize 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 presents compelling perceptual results but also significantly outperforms state-of-the-arts on the large-scale and challenging NIST IJB-A unconstrained face recognition benchmark. In addition, the proposed DA-GAN is also promising as a new approach for solving generic transfer learning problems more effectively. DA-GAN is the foundation of our submissions to NIST IJB-A 2017 face recognition competitions, where we won the 1st places on the tracks of verification and identification.

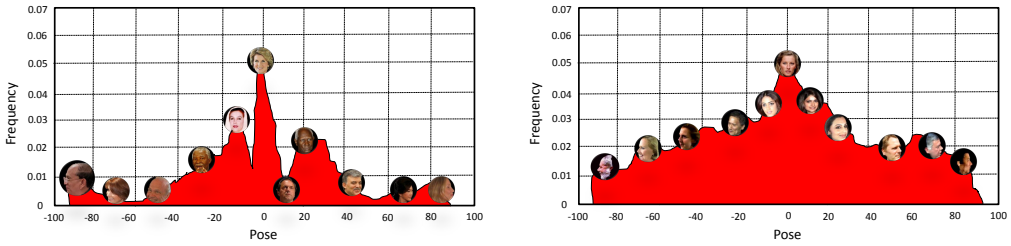
1 Introduction

Unconstrained face recognition is a very important yet extremely challenging problem. In recent years, deep learning techniques have significantly advanced large-scale unconstrained face recognition (8; 19; 27; 34; 29; 16), arguably driven by rapidly increasing resource of face images. However, labeling huge amount of data for feeding supervised deep learning algorithms is undoubtedly expensive and time-consuming. Moreover, as often observed in real-world scenarios, the pose distribution of available face recognition datasets (*e.g.*, IJB-A (15)) is usually unbalanced and has long-tail with

*Homepage: https://zhaoj9014.github.io/

†Jian Zhao and Zhecan Wang were interns at Panasonic R&D Center Singapore during this work.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



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

Figure 1: Comparison of pose distribution in the IJB-A dataset (15) w/o and w/ DA-GAN.

large pose variations, as shown in Figure. 1a. This has become a main obstacle for further pushing unconstrained face recognition performance. To address this critical issue, several research attempts (32; 31; 35) have been made to employ synthetic profile face images as augmented extra data to balance the pose variations.

However, naively learning from synthetic images can be problematic due to the distribution discrepancy between synthetic and real face images—synthetic data is often not realistic enough with artifacts and severe texture losses. The low-quality synthesis face images would mislead the learned face recognition model to overfit to fake information only presented in synthetic images and fail to generalize well on real faces. Brute-forcedly increasing the realism of the simulator is often expensive in terms of time cost and manpower, if possible.

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 focus on discriminating the realism of synthetic profile face images 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 pre-specified 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) (17) 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, DA-GAN introduces an auxiliary discriminative agent to enforce the generator to preserve identity information of the generated faces, which is critical for face recognition application. In addition, DA-GAN also imposes a pose perception loss to preserve pose and texture. The refined synthetic profile face images present 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 dual-agent GAN model, we impose a boundary equilibrium regularization term.

Experimental results show that DA-GAN not only presents compelling perceptual results but also 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) unconstrained face recognition benchmark (15). DA-GAN leads us to further win the 1st places on verification and identification tracks in the NIST IJB-A 2017 face recognition competitions. This strong evidence shows that our "recognition via generation" framework is effective and generic, and we expect that it benefits for more face recognition and transfer learning applications in the real world.

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 lost information inherent in projecting a 3D face into the 2D image space.
- We present qualitative and quantitative experiments showing the possibility of a "recognition via generation" framework and achieve the top performance on the challenging NIST IJB-A unconstrained face recognition benchmark (15) without extra human annotation efforts

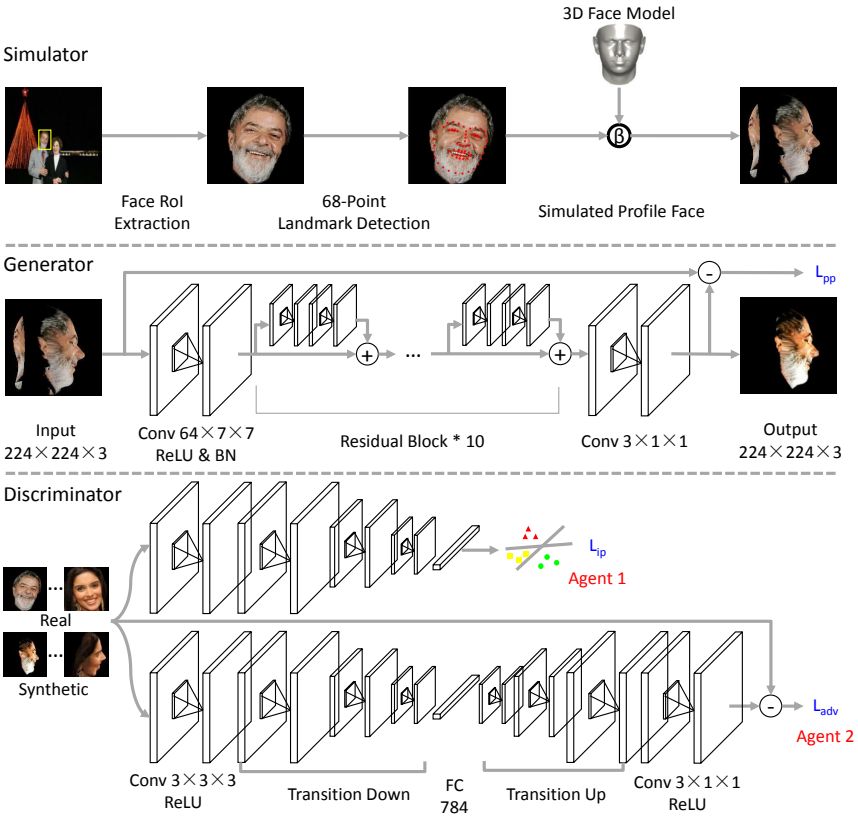


Figure 2: Overview of the proposed DA-GAN architecture. The simulator (upper panel) extracts face RoI, localizes landmark points and produces synthesis faces with arbitrary poses, which are fed to DA-GAN for realism refinement. DA-GAN uses a fully convolutional skip-net as the generator (middle panel) and an auto-encoder as the discriminator (bottom panel). The dual agents focus on both discriminating real *v.s.* fake (minimizing the loss L_{adv}) and preserving identity information (minimizing the loss L_{ip}). Best viewed in color.

by training deep neural networks on the refined face images together with real images. To our best knowledge, our proposed DA-GAN is the first model that is effective for automatically generating augmented data for face recognition in challenging conditions and indeed improves performance. DA-GAN won the 1st places on verification and identification tracks in the NIST IJB-A 2017 face recognition competitions.

2 Related works

As one of the most significant advancements on the research of deep generative models (14; 26), GAN has drawn substantial attention from the deep learning and computer vision community since it was first introduce by Goodfellow *et al.* (10). 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 (21) introduce the conditional version of GAN, to condition on both the generator and discriminator for effective image tagging. Berthelot *et al.* (2) propose a new **B**oundary **E**quilibrium **GAN** (BE-GAN) framework paired with a loss derived from the Wasserstein distance for training GAN, which derives a way of controlling the trade-off between image diversity and visual quality. These successful applications of GAN motivate us to develop profile view synthesis methods based on GAN. However, the generator of previous methods usually focus on generating 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, the 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, differnet from previous InfoGAN (5) which does not have the classification agent, and **A**uxiliary **C**lassifier **GAN** (AC-GAN) (22) which only performs classification, our propsoed DA-GAN performs face verification with an intrigued data augmentation. DA-GAN is a novel and practical model for efficient data augmentation and it is really effective in practice as proved in Sec. 4. DA-GAN generates the data in a completely different way from InfoGAN (5) and AC-GAN (22) which generate images from a random noise input or abstract semantic labels. Therefore, inferior to our model, those existing GAN-like models cannot exploit useful and rich prior information (*e.g.*, the shape, pose of faces) for effective data generation and augmentation. They cannot fully control the generated images. In contrast, DA-GAN can fully control the generated images and adjust the face pose (*e.g.*, yaw angles) distribution which is extremely unbalanced in real-world scenarios. 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 TP-GAN (13) that considers face synthesis based on GAN framework, and Apple GAN (28) that considers learning 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, whereas TP-GAN (13) tries to recover a frontal face from a profile view and Apple GAN (28) is designed for much simpler scenarios (*e.g.*, eye and hand image refinement); 2) TP-GAN (13) and Apple GAN (28) suffer from categorical information loss which limits their effectiveness in promoting recognition performance. In contrast, our proposed DA-GAN architecture effectively overcomes this issue by introducing dual discriminator agents.

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 is the main obstacle for learning a well-performed 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 Figure. 2, we first extracts the face **R**egion **o**f **I**nterest (RoI) from each available real face image, and estimate 68 facial landmark points using the **R**ecurrent **A**ttentive-**R**efinement (RAR) framework (31), which is robust to illumination changes and does not require a shape model in advance. We then estimate a transformation matrix between the detected 2D landmarks and the corresponding landmarks in the **3D M**orphable **M**odel (3D MM) using least-squares fit (35). 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 $\in \{[-90^o, -60^o] \cup [+60^o, +90^o]\}$) due to artifacts and severe texture losses, misleading the network to overfit to fake information only presented 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 truely benefical 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 (17; 9) and the leading performance of skip-net on recognition tasks (12; 33), we modify a skip-net (ResNet (12)) into a 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 are the network parameters for the generator, and H, W, and C denote the image height, width, and channel number, repectively.

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 maintains the contextual information, which is crucial for

artifact removal, fragement 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.

More 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:

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

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

$\mathcal{L}_{pp}$ 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:

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

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

Although $\mathcal{L}_{pp}$ may lead 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 the use of 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 $D_\phi : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{H \times W \times C}$ to be as simple as possible to avoid typical GAN tricks, which first projects the input real / fake face image into high-dimensional feature space through several **Conv**olution (Conv) and **F**ully **C**onnected (FC) layers of the encoder and then transformed back to the image-level representation through several **Deconv**olution (Deconv) and Conv layers of the decoder, as shown in Figure. 2. ϕ are the networks parameters for the discriminator. More details are provided in Sec. 4.

One agent of D_ϕ is trained with $\mathcal{L}_{adv}$ 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 (2),

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

where y denotes the real face image, k_t is a boundary equilibrium regularization term using Proportional Control Theory to maintain the equilibrium $\mathbb{E}[\sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)|] = \gamma \mathbb{E}[\sum_j |y_j - D_\phi(y_j)|]$, 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 thought of as a form of close-loop feedback control in which k_t is adjusted at each step.

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

The other agent of D_ϕ is trained with $\mathcal{L}_{ip}$ to preserve the identity discriminability of the refined face images. Specially, we define $\mathcal{L}_{ip}$ with the multi-class cross-entropy loss based on the output from the bottleneck layer of D_ϕ.

$$\begin{aligned}\mathcal{L}_{ip} = &\frac{1}{N}\sum_j -(Y_j log(D_\phi(y_j)) + (1-Y_j) log(1-D_\phi(y_j))) \\ &+ \frac{1}{N}\sum_i -(Y_i log(D_\phi(\tilde{x}_i)) + (1-Y_i) log(1-D_\phi(\tilde{x}_i))),\end{aligned} \tag{6}$$

where Y is the identity ground truth.

Thus, minimizing $\mathcal{L}_{ip}$ 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 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 $\mathcal{L}_{ip}$ alone makes the results prone to annoying artifacts, because the search for a local minimum of $\mathcal{L}_{ip}$ may go through a path that resides outside the manifold of natural face images. Thus, we combine $\mathcal{L}_{ip}$ with $\mathcal{L}_{adv}$ 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:

$$\mathcal{L}_{D_\phi} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{ip}. \tag{7}$$

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:

$$\begin{cases}\mathcal{L}_{D_\phi} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{ip}, \\ \mathcal{L}_{G_\theta} = (-\mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{ip}) + \lambda_2 \mathcal{L}_{pp}.\end{cases} \tag{8}$$

We optimize DA-GAN by alternatively optimizing D_ϕ and G_θ for each training iteration. Similar as in (2), 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:

$$\mathcal{L}_{con} = \sum_j |y_j - D_\phi(y_j)| + |\gamma \sum_j |y_j - D_\phi(y_j)| - \sum_i |\tilde{x}_i - D_\phi(\tilde{x}_i)||. \tag{9}$$

$\mathcal{L}_{con}$ can be used to determine when the network has reached its final state or if the model has collapsed. Detailed algorithm on the training procedures is provided in supplementary material Sec. 1.

4 Experiments

4.1 Experimental settings

Benchmark dataset: 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. Therefore, we evaluate the possibility of "recognition via generation" of DA-GAN on the most challenging unconstrained face recognition benchmark dataset IJB-A (15). IJB-A (15) contains both images and video frames from 500 subjects with 5,397 images and 2,042 videos that are split into 20,412 frames, 11.4 images and 4.2 videos per subject, captured from in-the-wild environment to avoid the near frontal bias, 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, respectively. More details are provided in supplementary material Sec. 2.

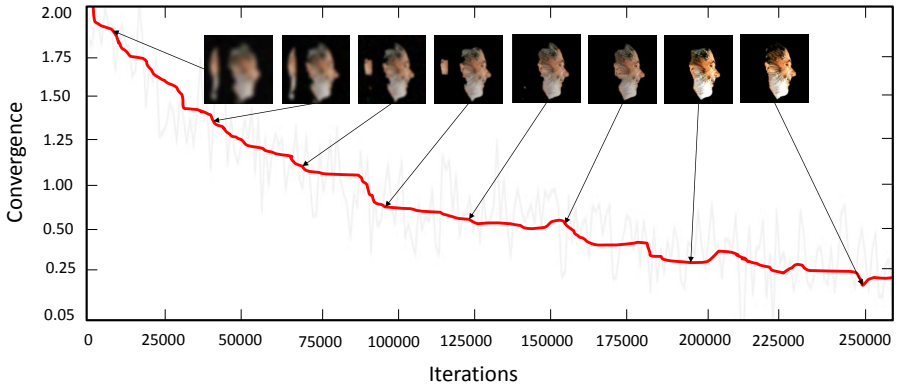
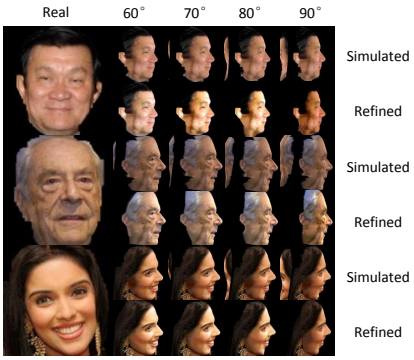
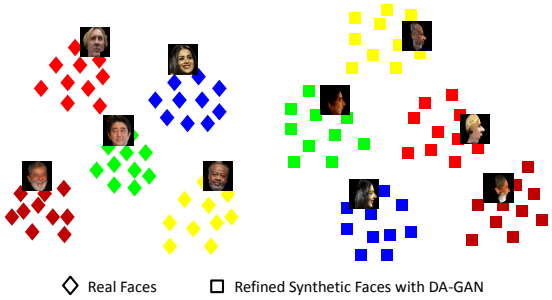


Figure 3: Quality of refined results *w.r.t.* the network convergence measurement $\mathcal{L}_{con}$.



(a) Refined results of DA-GAN.



(b) Feature space of real faces and DA-GAN synthetic faces.

Figure 4: Qualitative analysis of DA-GAN.

Reproducibility: The proposed method is implementated by extending the Keras framework (6). All networks are trained on three NVIDIA GeForce GTX TITAN X GPUs with 12GB memory for each. Please refer to supplementary material Sec. 3 & 4 for full details on network architectures and training procedures.

4.2 Results and discussions

Qualitative results – 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 $\mathcal{L}_{con}$, as shown in Figure. 3. As can be seen, our DA-GAN ensures a fast yet stable convergence through the carefully designed optimization scheme and boundary equilibrium regularization term. The network convergence measurement $\mathcal{L}_{con}$ correlates well with image fidelity.

Most of the previous works (31; 32; 35) on profile view synthesis are dedicated to address this problem within a pose range of $\pm 60^o$. Because it is commonly believed that with a pose that is larger than 60^o, it is difficult for a model to generate faithful profile view images. Similarly, our simulator is also good at normalizing small posed faces while suffers severe artifacts and texture losses under large poses (*e.g.*, yaw angles $\in \{[-90^o, -60^o] \cup [+60^o, +90^o]\}$), as shown in Figure. 4a the first row 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 Figure. 4a the second row for each subject. Compared with the raw simulated faces, the refined results by DA-GAN present a good photorealistic quality. More visualized samples are provided in supplementary material Sec. 5.

To verify the superiority of DA-GAN as well as the contribution of each component, we also compare the qualitative results produced by the vanilla GAN (10), Apple GAN (28), BE-GAN (2) and three variations of DA-GAN in terms of w/o $\mathcal{L}_{adv}$, $\mathcal{L}_{ip}$, $\mathcal{L}_{pp}$ in each case, repectively. Please refer to supplementary material Sec. 5 for details.

Table 1: Performance comparison of DA-GAN with state-of-the-arts on IJB-A 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	**Face verification**		
	TAR @ FAR=0.10	**TAR @ FAR=0.01**	**TAR @ FAR=0.001**
OpenBR (15)	0.433 ± 0.006	0.236 ± 0.009	0.104 ± 0.014
GOTS (15)	0.627 ± 0.012	0.406 ± 0.014	0.198 ± 0.008
Pooling faces (11)	0.631	0.309	-
LSFS (30)	0.895 ± 0.013	0.733 ± 0.034	0.514 ± 0.060
Deep Multi-pose (1)	0.911	0.787	-
DCNN$_{manual}$ (4)	0.947 ± 0.011	0.787 ± 0.043	-
Triplet Similarity (27)	0.945 ± 0.002	0.790 ± 0.030	0.590 ± 0.050
VGG-Face (23)	-	0.805 ± 0.030	-
PAMs (19)	0.652 ± 0.037	0.826 ± 0.018	-
DCNN$_{fusion}$ (3)	0.967 ± 0.009	0.838 ± 0.042	-
Masi *et al.* (20)	-	0.886	0.725
Triplet Embedding (27)	0.964 ± 0.005	0.900 ± 0.010	0.813 ± 0.020
All-In-One (25)	0.976 ± 0.004	0.922 ± 0.010	0.823 ± 0.020
Template Adaptation (8)	0.979 ± 0.004	0.939 ± 0.013	0.836 ± 0.027
NAN (34)	0.978 ± 0.003	0.941 ± 0.008	0.881 ± 0.011
L_2-softmax (24)	0.984 ± 0.002	0.970 ± 0.004	0.943 ± 0.005
b-1	0.989 ± 0.003	0.963 ± 0.007	0.920 ± 0.006
b-2	0.978 ± 0.003	0.950 ± 0.009	0.901 ± 0.008
DA-GAN (ours)	$\mathbf{0.991 \pm 0.003}$	$\mathbf{0.976 \pm 0.007}$	$\mathbf{0.930 \pm 0.005}$

Table 2: Performance comparison of DA-GAN with state-of-the-arts on IJB-A identification protocol. For FNIR metric, a lower number means better performance. For the other metrics, a higher number means better performance. The results offered by our proposed method are highlighted in bold.

Method	**Face identification**			
	FNIR @ FPIR=0.10	**FNIR @ FPIR=0.01**	**Rank1**	**Rank5**
OpenBR (15)	0.851 ± 0.028	0.934 ± 0.017	0.246 ± 0.011	0.375 ± 0.008
GOTS (15)	0.765 ± 0.033	0.953 ± 0.024	0.433 ± 0.021	0.595 ± 0.020
B-CNN (7)	0.659 ± 0.032	0.857 ± 0.027	0.588 ± 0.020	0.796 ± 0.017
LSFS (30)	0.387 ± 0.032	0.617 ± 0.063	0.820 ± 0.024	0.929 ± 0.013
Pooling faces (11)	-	-	0.846	0.933
Deep Multi-pose (1)	0.250	0.480	0.846	0.927
DCNN$_{manual}$ (4)	-	-	0.852 ± 0.018	0.937 ± 0.010
Triplet Similarity (27)	0.246 ± 0.014	0.444 ± 0.065	0.880 ± 0.015	0.950 ± 0.007
VGG-Face (23)	0.33 ± 0.031	0.539 ± 0.077	0.913 ± 0.011	-
PAMs (19)	-	-	0.840 ± 0.012	0.925 ± 0.008
DCNN$_{fusion}$ (3)	0.210 ± 0.033	0.423 ± 0.094	0.903 ± 0.012	0.965 ± 0.008
Masi *et al.* (20)	-	-	0.906	0.962
Triplet Embedding (27)	0.137 ± 0.014	0.247 ± 0.030	0.932 ± 0.010	-
Template Adaptation (8)	0.118 ± 0.016	0.226 ± 0.049	0.928 ± 0.010	0.977 ± 0.004
All-In-One (25)	0.113 ± 0.014	0.208 ± 0.020	0.947 ± 0.008	-
NAN (34)	0.083 ± 0.009	0.183 ± 0.041	0.958 ± 0.005	0.980 ± 0.005
L_2-softmax (24)	0.044 ± 0.006	0.085 ± 0.041	0.973 ± 0.005	-
b-1	0.068 ± 0.010	0.125 ± 0.035	0.966 ± 0.006	0.987 ± 0.003
b-2	0.108 ± 0.008	0.179 ± 0.042	0.960 ± 0.007	0.982 ± 0.004
DA-GAN (ours)	$\mathbf{0.051 \pm 0.009}$	$\mathbf{0.110 \pm 0.039}$	$\mathbf{0.971 \pm 0.007}$	$\mathbf{0.989 \pm 0.003}$

To gain insights into the effectivenss of identity preserving quality of our DA-GAN, we further use t-SNE (18) to visualize the deep features of both refined profile faces and real faces in a 2D space in Figure. 4b. As can be seen, 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 $\mathcal{L}_{ip}$.

Quantitative results – "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 benchmark dataset (15) with three different settings. In the three settings,

the pre-trained deep recognition models are respectively fine-tuned on the original training data of each split without extra data (baseline 1: b-1), the original training data of each split with extra synthetic faces by our simulator (baseline 2: b-2), and the original training data of each split with extra refined faces by our DA-GAN (our method: "recognition via generation" framework based on DA-GAN, DA-GAN for short). The performance comparison of DA-GAN with the two baselines and other state-of-the-arts on IJB-A (15) unconstrained face verification and identification protocols are given in Table. 1 and Table. 2.

We can observe that even with extra training data, b-2 presents inferior performance than b-1 for all metrics of both face verification and identification. This demonstrates that naively learning from synthetic images can be problematic due to a gap between synthetic and real image distributions – synthetic data is often not realistic enough with artifacts and severe texture losses, misleading the network to overfit to fake information only presented 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 method outperforms b-1 by 1.00% for TAR @ FAR=0.001 of verification and 1.50% for FNIR @ FPIR=0.01, 0.50% for Rank-1 of identification. Our method achieves comparable performance with L_2-softmax (24), which employ a much more computational complex recognition model even without fine-tuning or template adaptation procedures as we do. Moreover, DA-GAN outperforms NAN (34) by 4.90% for TAR @ FAR=0.001 of verification and 7.30% for FNIR @ FPIR=0.01, 1.30% for Rank-1 of identification. These results won the 1st places on verification and identification tracks in NIST IJB-A 2017 face recognition competitions[3]. This well verified the promissing 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 (15) split1 to gain insights into unconstrained face recognition with the proposed "recognition via generation" framework based on DA-GAN. For fully detailed visualization results in high resolution and corresponding analysis, please refer to supplementary material Sec. 6 & 7.

5 Conclusion

We propose 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 lost information inherent 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 promissing 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 unconstrained face recognition benchmark without extra human annotation efforts. Based on DA-GAN, we won the 1st places on verification and identification tracks in NIST IJB-A 2017 face recognition competitions. It would be interesting to apply DA-GAN for other transfer learning applications in the future.

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133, Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112 and NUS IDS grant R-263-000-C67-646.

We would like to thank Junliang Xing (Institute of Automation, Chinese Academy of Sciences), Hengzhu Liu, Xucan Chen, and Yongping Zhai (National University of Defense Technology) for helpful discussions.

[3]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.

References

[1] 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 *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[4] 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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 118–126, 2015.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[6] F. Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[7] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller. One-to-many face recognition with bilinear cnns. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[8] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[9] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–67, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] 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. *arXiv preprint arXiv:1704.04086*, 2017.

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] 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 *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[16] 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 *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1068–1072. ACM, 2016.

[17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[18] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[19] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, 2016.

[20] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? *arXiv preprint arXiv:1603.07057*, 2016.

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition.

[24] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[25] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv preprint arXiv:1611.00851*, 2016.

[26] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[27] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–8. IEEE, 2016.

[28] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[30] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015.
[31] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision*, pages 57–72. Springer, 2016.
[32] S. Xiao, L. Liu, X. Nie, J. Feng, A. A. Kassim, and S. Yan. A live face swapper. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 691–692. ACM, 2016.
[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.
[34] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016.
[35] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3d morphable model fitting. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

Supplementary Material for Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

Jian Zhao[1,2*†] **Lin Xiong**[3] **Karlekar Jayashree**[3] **Jianshu Li**[1] **Fang Zhao**[1]
Zhecan Wang[4†] **Sugiri Pranata**[3] **Shengmei Shen**[3]
Shuicheng Yan[1,5] **Jiashi Feng**[1]
[1]National University of Singapore [2]National University of Defense Technology
[3] Panasonic R&D Center Singapore [4] Franklin. W. Olin College of Engineering
[5] Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu {lin.xiong, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com
zhecan.wang@students.olin.edu {elezhf, eleyans, elefjia}@u.nus.edu

Abstract

In this supplementary material, we present fully detailed information on 1) learning algorithm of the proposed **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (DA-GAN) model; 2) details on the IJB-A benchmark dataset (6); 3) network architectures; 4) training details; 5) qualitative analysis of DA-GAN; 6) high-resolution visualized verification results for IJB-A (6) split1; 7) high-resolution visualized identification results for IJB-A (6) split1.

1 Learning algorithm of DA-GAN model

We summarize detailed the training procedures of our DA-GAN in Algorithm. 1.

2 Details on the IJB-A benchmark dataset

IJB-A (6) contains both images and video frames from 500 subjects with 5,397 images and 2,042 videos that are split into 20,412 frames, 11.4 images and 4.2 videos per subject, captured from in-the-wild environment to avoid the near frontal bias, 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, respectively.

IJB-A (6) 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 the subject matching a probe identity 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,

*Homepage: https://zhaoj9014.github.io/

†Jian Zhao and Zhecan Wang were interns at Panasonic R&D Center Singapore during this work.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Algorithm 1 Learning algorithm of DA-GAN

Input: Sets of synthetic profile face images x_i, real face images y_j, and the associated identity labels Y_i, max number of epoches (nb_e), batch size (b), number of network updates per step (nb_s), input size (im_w, im_h, im_c), weight decay, learning rate (lr), k_0, λ_1, λ_2, α, γ;

Output: DA-GAN generator G_θ and discriminator D_ϕ;

for e=1, $\cdots$, nb_e **do**

  for s=1, $\cdots$, nb_s **do**

    1. Optimize D_ϕ;
    2. Optimize G_θ;
    3. Update k_t;
    4. Measure network convergence $\mathcal{L}_{con}$;
    5. Visualize intermediate results;

  end for

  Archive G_θ and D_ϕ models for each training epoch;

end for

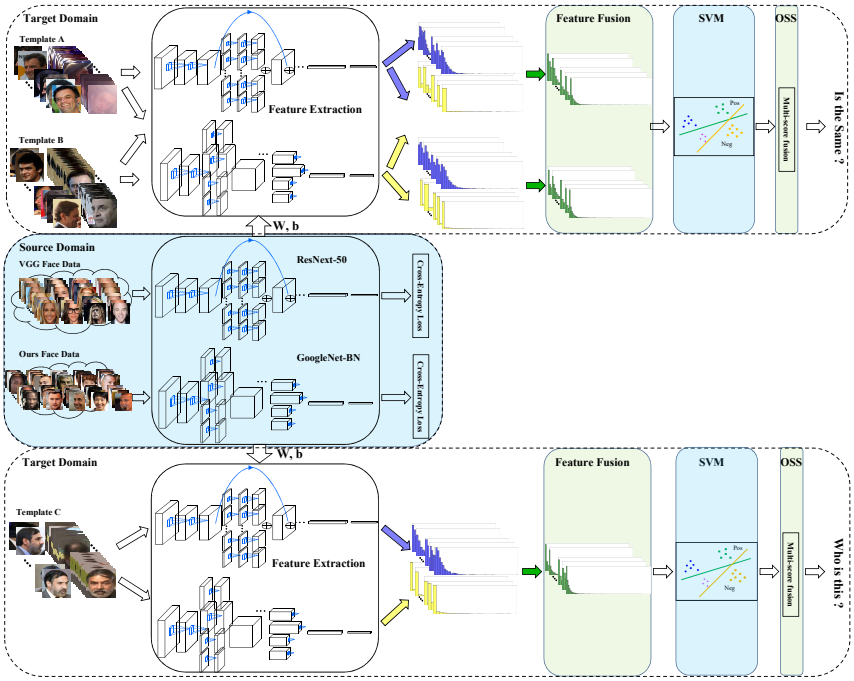


Figure 1: Framework overview of "recognition via generation". We transfer learn two state-of-the-art deep neural networks – ResNext-50 (10) and GoogleNet-BN (8) from source domain to target domain extended by DA-GAN. We ensemble the compensate two-view information from the two models to train template adapted SVMs (2). The resulted margins are robust and discriminative for unconstrained face recognition. Best viewed in color.

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 that 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 (6).

3 Network architectures

- Simulator: RAR framework (9) (face RoI extraction & 68 facial landmark detection), 3D MM (11) (profile face image simulation with pre-defined yaw angles).
- Generator: Input $224 \times 224 \times 3$, Conv $64 \times 7 \times 7$, ReLU[3], BN[4], $10*$Residual block (Conv $64 \times 7 \times 7$, ReLU, BN, Conv $64 \times 7 \times 7$, Ele-Sum[5], ReLU, BN), Conv $3 \times 1 \times 1$.

[3]ReLU is short for Rectified Linear Units (4).

[4]BN is short for Batch Normalization (5).

[5]Ele-Sum is short for element-wise summation.

- 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 $256 \times 3 \times 3/2$, ReLU, Conv $384 \times 3 \times 3$, 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 $224 \times 224 \times 3$, ResNext-50 (cardinality $= 32$) (10) & GoogleNet-BN (8) (model fusion), template adapted **S**upport **V**ector **M**achine (SVM) (2) (metric learning).

The overview of our proposed "recognition via generation" framework is illustrated in Figure. 1. We transfer learn two state-of-the-art deep neural networks – ResNext-50 (10) and GoogleNet-BN (8) from source domain (MS-Celeb-1M (3), removed overlapping parts with IJB-A (6)) to target domain of IJB-A (6) extended by DA-GAN. We ensemble the compensate two-view information (learned deep features) from the ResNext-50 (10) and GoogleNet-BN (8) models to train template adapted SVMs (2). The resulted margins are robust and discriminative for unconstrained face recognition.

4 Training details

- DA-GAN: 1) Extract face RoIs from the available training data of each IJB-A (6) split, and detect 68 facial landmark points using the RAR framework (9). 2) Simulate profile faces with pre-defined yaw angles $\in \{\pm 10, \pm 20, \pm 30, \pm 40, \pm 50, \pm 60, \pm 70, \pm 80, \pm 90\}$ using 3D MM (11). 3) Train DA-GAN using Adam with mini-batch (FC 333 with Softmax appended to the output of the bottleneck layer of D_ϕ for $\mathcal{L}_{ip}$ 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 $\mathcal{L}_{con}$ 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 and divided by 10 every 30 epoches; set the weight decay to 1×10^{-4}; set the momentum to 0.9. 2) Pre-process MS-Celeb-1M (3) data, including overlapping part removal with IJB-A (6) and face RoI extraction, resulting in $4{,}356{,}052$ face images for $53{,}317$ subjects in total. 3) Train ResNext-50 (cardinality $= 32$) (10) & GoogleNet-BN (8) using **S**tochastic **G**radient **D**escent (SGD) on the cleaned MS-Celeb-1M (3) data. 4) Reset the learning rate to 0.0001 and divided by 10 every 10 epoches. 5) Inject the refined profile face images and video frames into IJB-A (6) each split training data and fine-tune the pre-trained deep recognition models.

- Template adapted SVM models: 1) Concat the learned pose-invariant features from the penultimate layers of deep recognition models ($\mathbb{R}^{2048}$ C-Sum[6] $\mathbb{R}^{1024} \mapsto \mathbb{R}^{3072}$). 2) Train template adapted SVM models similarly as introduced in (2).

More formally, the template adapted SVMs are learned by optimizing the following L_2-regularized objective function:

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

where $f_F(\cdot)$ denotes the non-linear function learned by our deep recognition models, x denote the face media, w denote the weights including bias term, $y_i \in \{-1, 1\}$ denotes the label indicating whether the current sample being negative or possible, N_+ indicates the number of positive samples, N_- indicates the number of negative ones, $N_- \gg N_+$, the constraint for negative samples $\lambda_- = C\frac{N_+ + N_-}{2N_-}$, the constraint for positive samples $\lambda_+ = C\frac{N_+ + N_-}{2N_+}$, C is a trade-off factor, and we set it to 20 in our method.

[6] C-Sum is short for concat.

Since a template contains both face images and / or video frames, containing large variances in terms of media modality, pose, expression, occlusion, and illumination. In order to better address the underlying distracting factors within each template, we split each template into several sub-templates 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.

Let t_j^V be the mean encoding of the jth video sequence, then

$$t_j^V = \frac{1}{N_j^V} \sum_{i=1}^{N_j^V} f_F(\mathbf{x}_i), \tag{2}$$

where N_j^V is the number of frame in the jth video sequence, $\mathbf{x}_i$ denotes the ith frame of video j.

Thus, the representations for the ath template can be expressed as

$$T_a = \left\{ t_i^I, ..., t_{N_a}^V \right\}, \tag{3}$$

where t_i^I denotes the sub-template for the ith image, $t_{N_a}^V$ 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 (2). For verification, the positive sample of template specific SVM is a probe template, and the 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; whereas for gallery template specific SVM, other gallery templates and the whole training set are bundled together as the large-scale negative samples.

Based on one shot similarity, we compute the fine-grained similarity between two sub-template representations p and q via $s(p,q) = \frac{1}{2}\mathcal{P}(q) + \frac{1}{2}\mathcal{Q}(p)$, 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. (3), a template may contain various number 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,

$$s(T_a, T_b) = \frac{\sum\limits_{t_i \in T_a, t_j \in T_b} s(t_i, t_j)\, e^{\beta\, s(t_i, t_j)}}{\sum\limits_{t_i \in T_a, t_j \in T_b} e^{\beta\, s(t_i, t_j)}}, \tag{4}$$

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

5 Qualitative analysis of DA-GAN

We visualize the high-resolution refined results of DA-GAN under various poses with yaw angles ranging from -90^o to -10^o and $+10^o$ to $+90^o$ at a stride of 10^o in Figure. 2 and Figure. 3 to verify the compelling perceptual quality of DA-GAN. As can be seen, DA-GAN is able to adaptively remove artifacts (*e.g.*, face fragments and black holes) introduced by the simulator, stitch fragments, and compensate texture losses in terms of facial details and color realism, especially for large poses. As a result, the refined faces of DA-GAN present more intuitively photorealistic and natural characteristics.

To verify the superiority of DA-GAN as well as the contribution of each component, we also compare the qualitative results produced by the vanilla GAN, Apple GAN (7), BE-GAN (1), and three variations of DA-GAN in terms of w/o $\mathcal{L}_{adv}$, $\mathcal{L}_{ip}$, $\mathcal{L}_{pp}$ in each case, repectively. As shown in Figure. 4, inference without $\mathcal{L}_{ip}$ deviates from the true appearance seriously, and the synthesis without $\mathcal{L}_{adv}$ tends to be very blurry, while the results without the $\mathcal{L}_{pp}$ sometimes show blurry and unnatural effect with strange artifacts / color involved. Compared with vanilla GAN, Apple GAN (7) and BE-GAN (1), which all fail with poses larger than 60^o, our DA-GAN presents a good identity preserving quality while producing photorealistic synthesis.

6 Verification result analysis for IJB-A Split1

For face verification, after computing the similarities for all pairs of probe and reference sets, we sort the resulting list. Each row represents a probe and reference template pair. The original templates

within IJB-A (6) 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. Figure. 5 shows the best matched cases. In the top-30 scoring correct matches, we immediately note that every reference template contains dozens of media. The probe templates either contain dozens of media or one medium that matches well. Figure. 6 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. Figure. 7 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. These extremely difficult cases cannot be solved even using the specific operations designed in our "recognition via generation" framework. Figure. 8 illustrating the worst non-matched cases highlights the understandable errors, representing impostors in challenging modalities.

7 Identification result analysis for IJB-A Split1

For face identification, Figure. 9 $1st$-column shows the query images from probe templates. Figure. 9 column 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 performs successful searching in Rank-1, which well proved the effectiveness of our DA-GAN based method for generic transfer learning and face-centric analysis. It would be interesting to apply DA-GAN for other transfer learning applications in the future.

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133, Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112 and NUS IDS grant R-263-000-C67-646.

We would like to thank Junliang Xing (Institute of Automation, Chinese Academy of Sciences), Hengzhu Liu, Xucan Chen, and Yongping Zhai (National University of Defense Technology) for helpful discussions.

References

[1] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[2] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[3] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] 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 *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[7] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[9] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision*, pages 57–72. Springer, 2016.

[10] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[11] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3d morphable model fitting. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

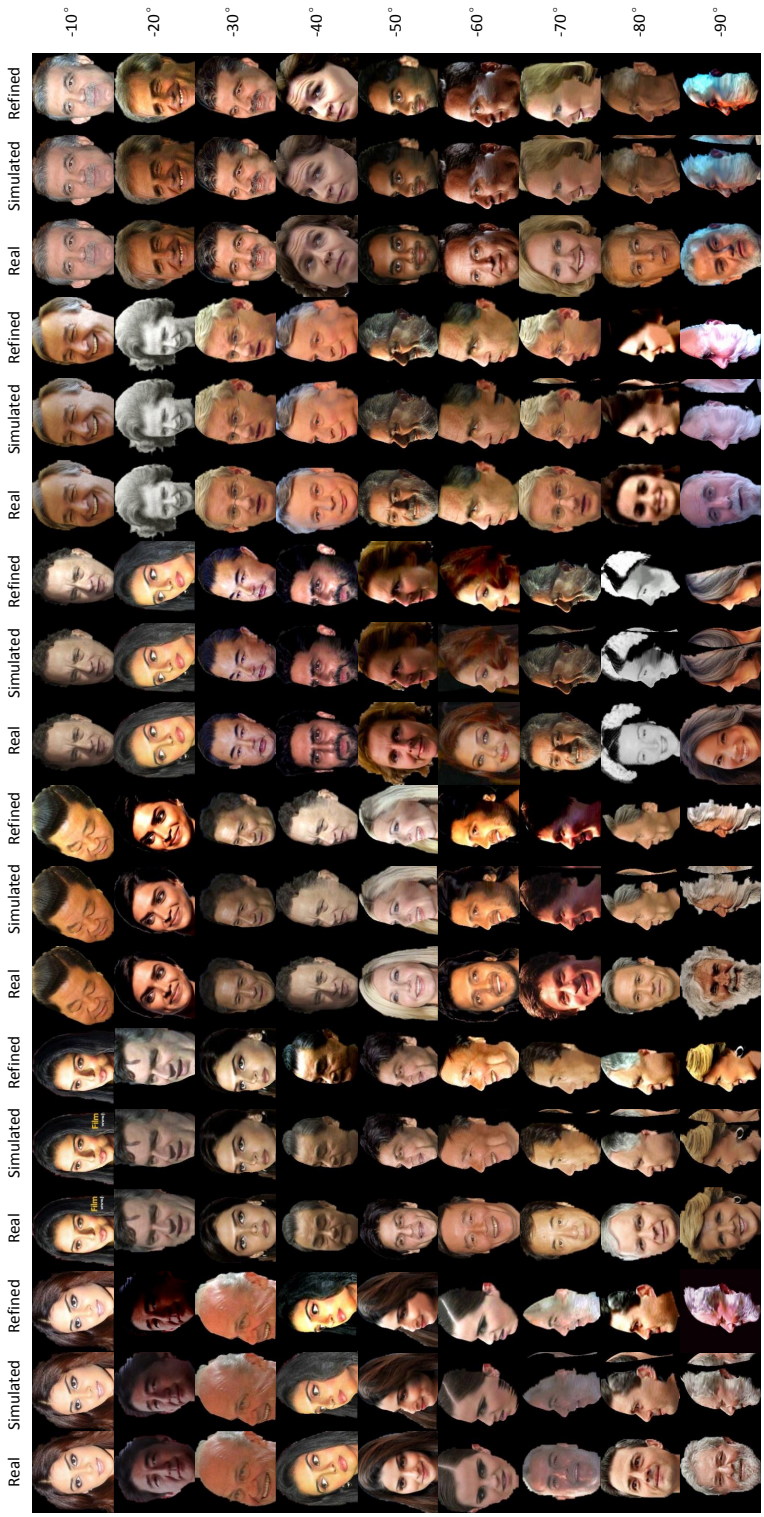


Figure 2: Refined results of DA-GAN under various poses with yaw angles ranging from -90^o to -10^o at a stride of 10^o.

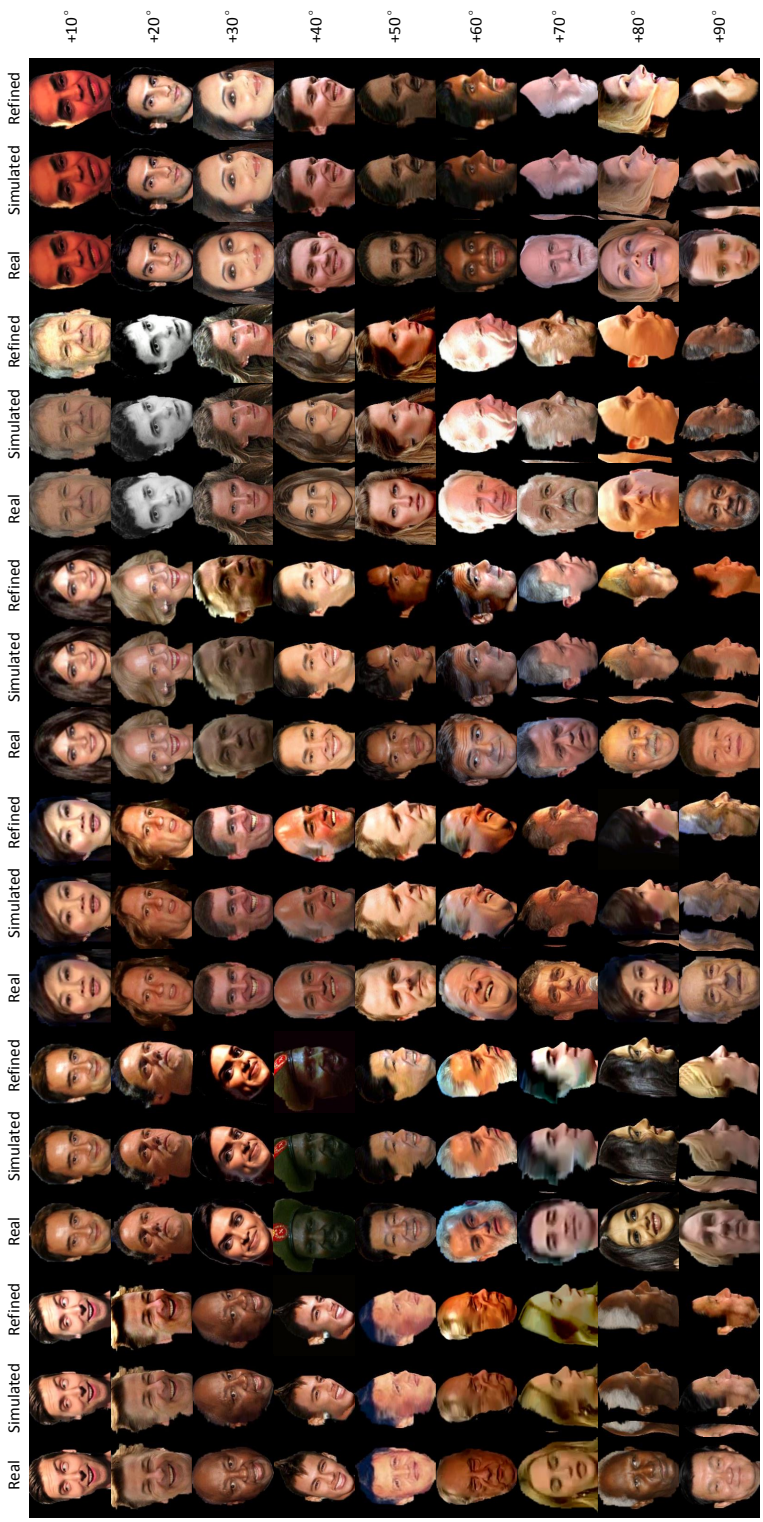


Figure 3: Refined results of DA-GAN under various poses with yaw angles ranging from $+10^o$ to $+90^o$ at a stride of 10^o.

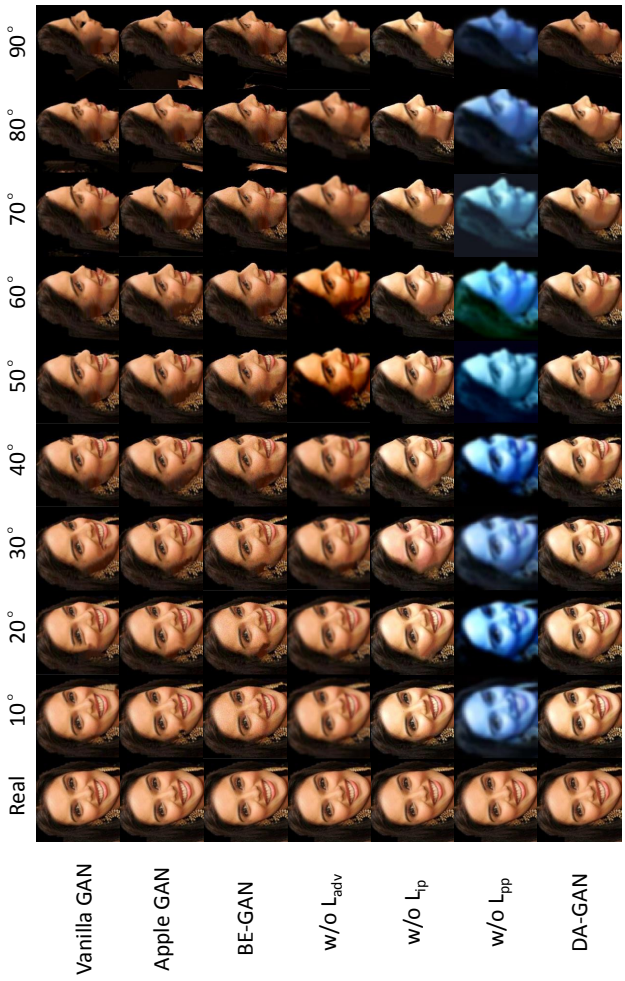


Figure 4: Qualitative results comparison of DA-GAN with state-of-the-art GANs and three different network settings.

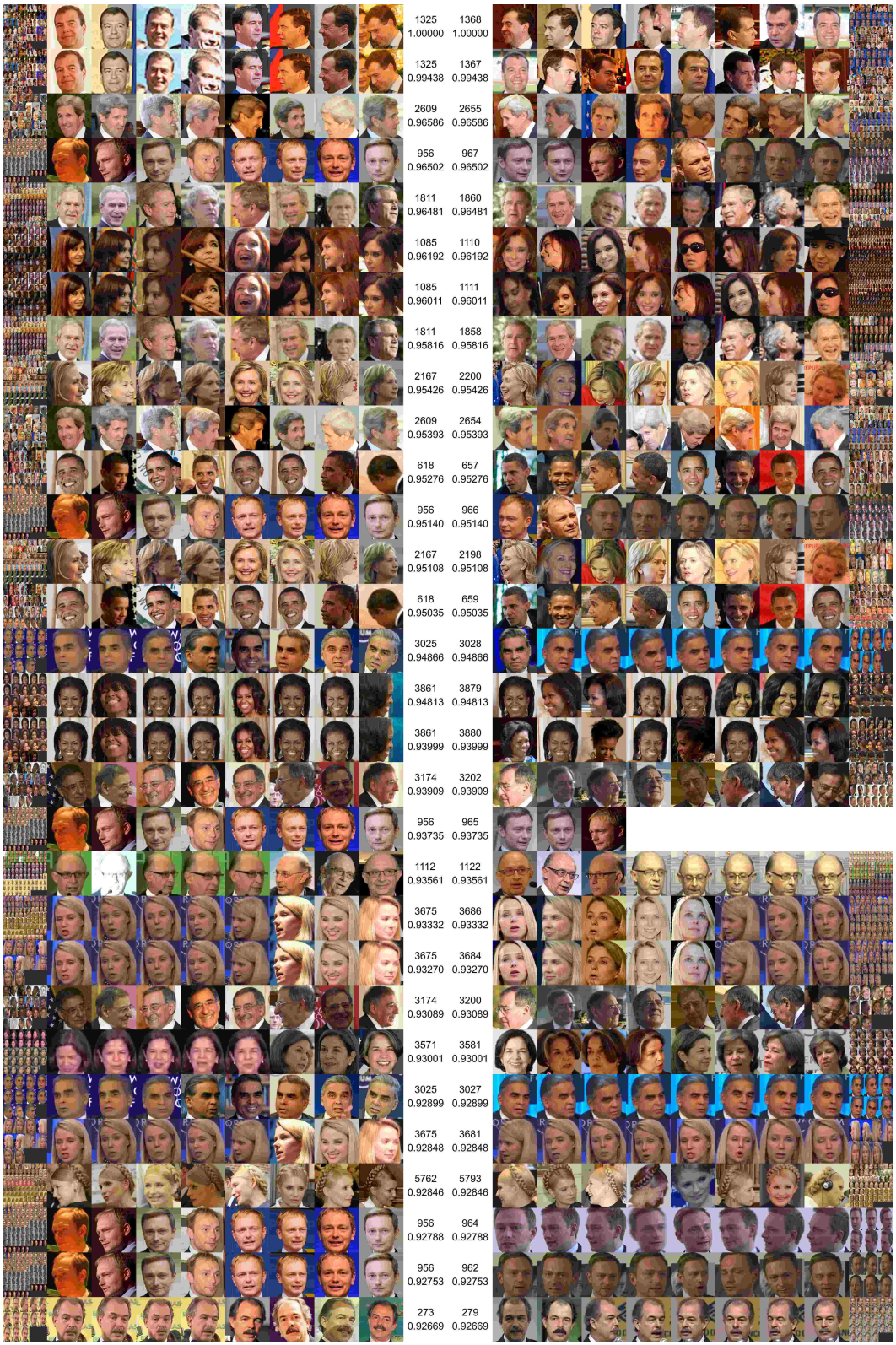


Figure 5: Verification results analysis for best matched cases on IJB-A split1.

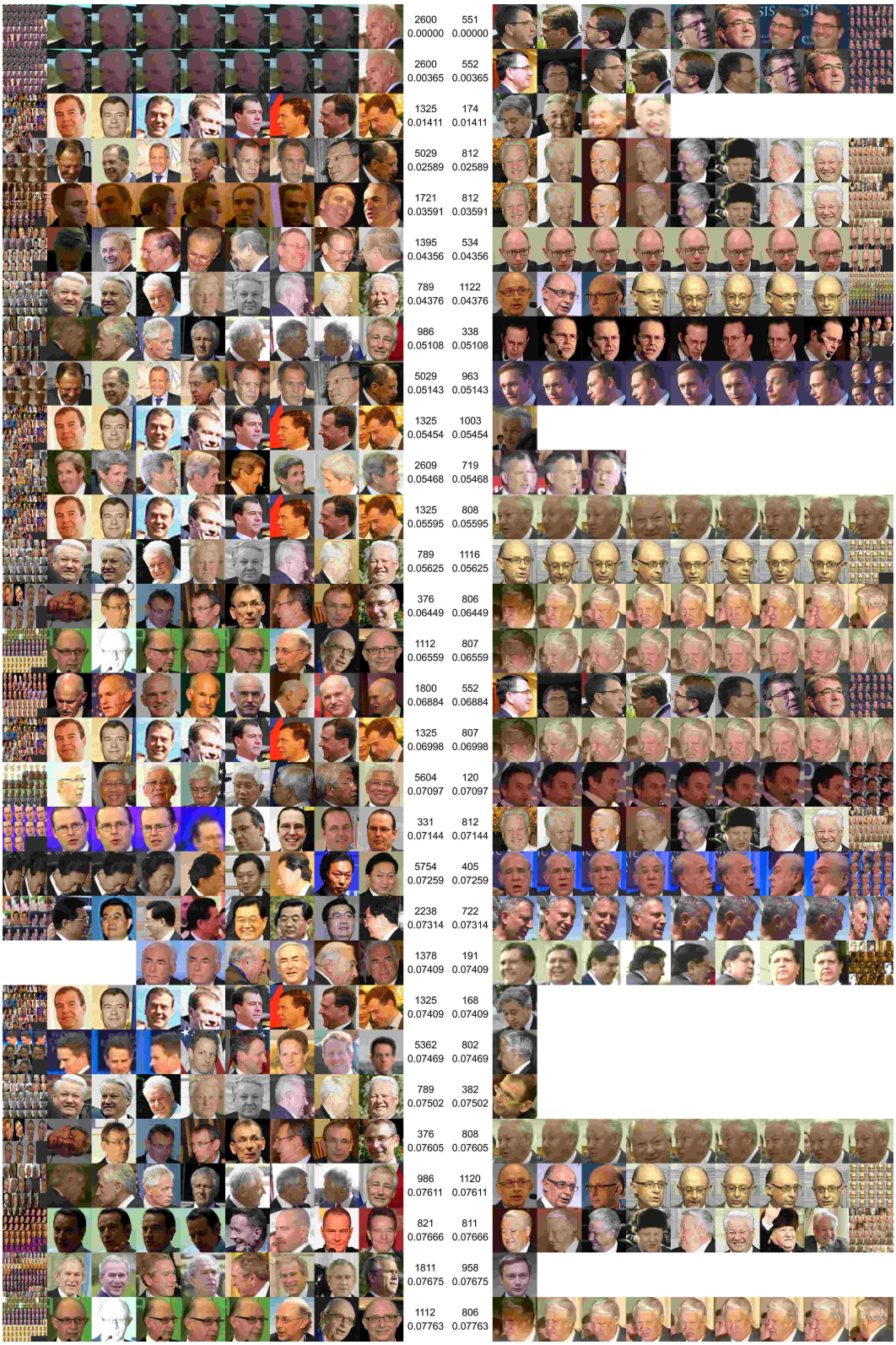


Figure 6: Verification results analysis for best non-matched cases on IJB-A split1.

Figure 7: Verification results analysis for worst matched cases on IJB-A split1.

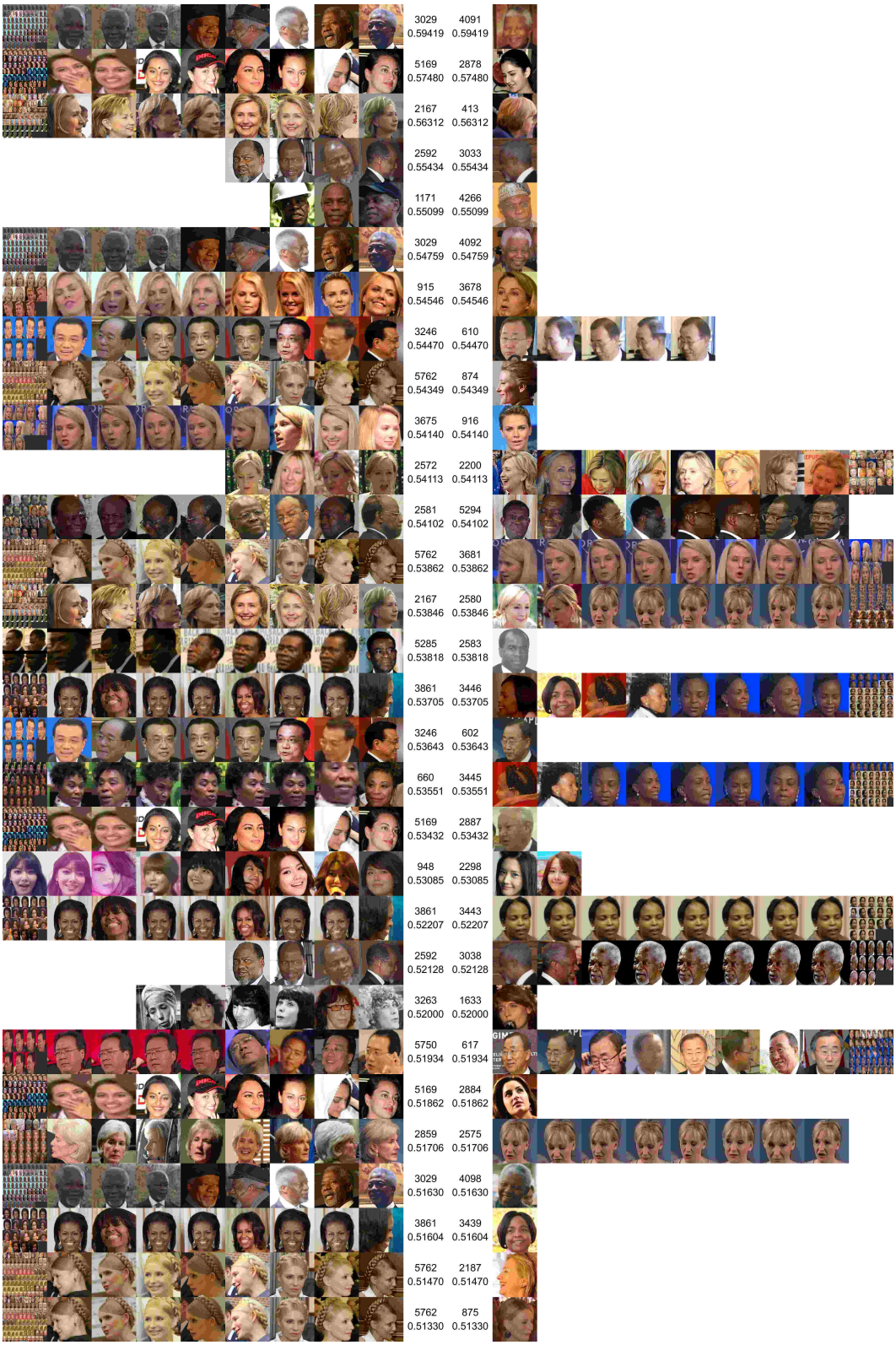


Figure 8: Verification results analysis for worst non-matched cases on IJB-A split1.

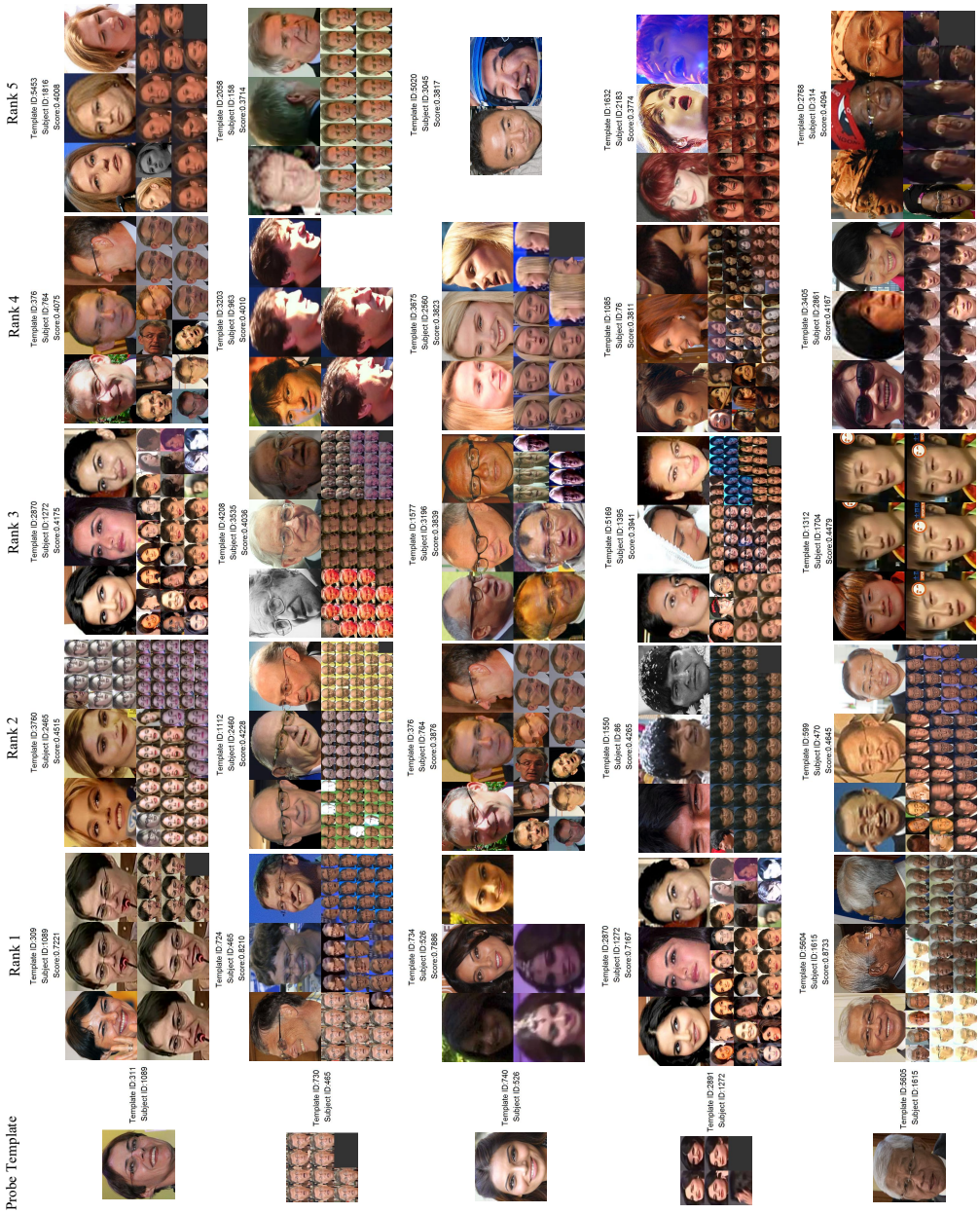


Figure 9: Identification results analysis on IJB-A split1.