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

Jian Zhao^{1,2*} Lin Xiong³ Karlekar Jayashree³ Jianshu Li¹ Fang Zhao¹

Zhecan Wang^{4†} Sugiri Pranata³ Shengmei Shen³

Shuicheng Yan^{1,5} Jiashi Feng¹

¹National University of Singapore ²National University of Defense Technology

³Panasonic R&D Center Singapore ⁴Franklin. W. Olin College of Engineering

⁵Qihoo 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 Dual-Agent Generative Adversarial Network (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 **Receiver Operating Characteristic (ROC)** analysis measures the **True Accept Rate (TAR)**, which is the fraction of genuine comparisons that correctly exceed the threshold, and the **False Accept Rate (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 **Cumulative Match Characteristic (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.

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), $k_0, \lambda_1, \lambda_2, \alpha, \gamma$;

Output: DA-GAN generator G_θ and discriminator D_ϕ ;

```

for e=1, ..., nb_e do
    for s=1, ..., nb_s do
        1. Optimize  $D_\phi$ ;
        2. Optimize  $G_\theta$ ;
        3. Update  $k_t$ ;
        4. Measure network convergence  $\mathcal{L}_{con}$ ;
        5. Visualize intermediate results;
    end for
    Archive  $G_\theta$  and  $D_\phi$  models for each training epoch;
end for

```

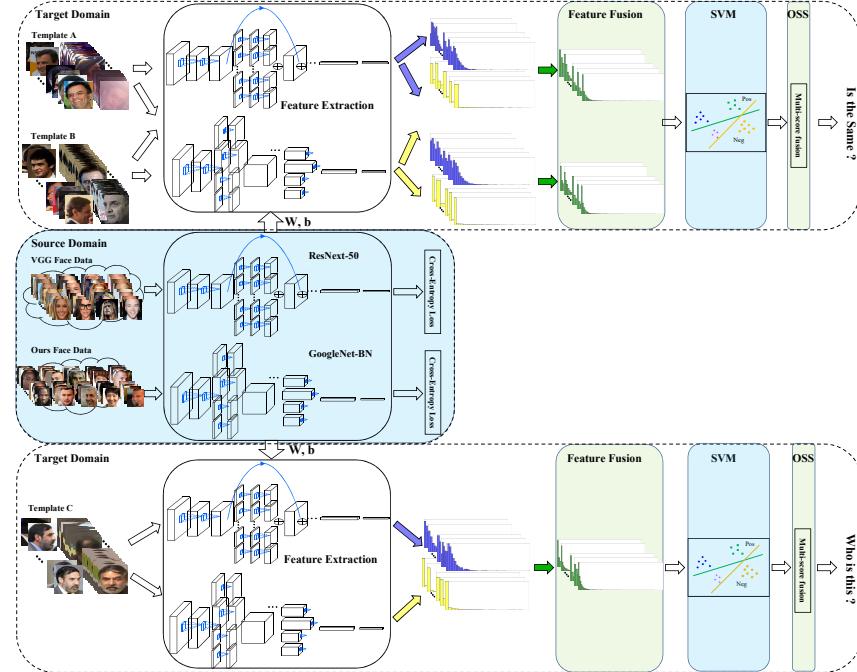


Figure 1: Framework overview of “recognition via generation”. We transfer learn two state-of-the-art deep neural networks – ResNext-50 (10) and GoogleNet-BN (8) from source domain to target domain extended by DA-GAN. We ensemble the compensate two-view information from the two models to train template 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 **False Positive Identification Rate** (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 **False Negative Identification Rate** (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 $64 \times 7 \times 7$, ReLU³, BN⁴, 10*Residual block (Conv $64 \times 7 \times 7$, ReLU, BN, Conv $64 \times 7 \times 7$, Ele-Sum⁵, ReLU, BN), Conv $3 \times 1 \times 1$.

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

⁴BN is short for Batch Normalization (5).

⁵Ele-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 Support Vector Machine (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) Select out near frontal face images and video frames from IJB-A (6) each split training data based on the criterion of yaw angle $\leq 10^\circ$ && forehead visible && eyes visible && nose and mouth visible. 2) Extract face RoIs and detect 68 facial landmark points using the RAR framework (9). 3) 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). 4) 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 Stochastic Gradient Descent (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-Sum⁶ $\mathbb{R}^{1024} \mapsto \mathbb{R}^{3072}$). 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:

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

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 positive, 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.

⁶C-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 j th video sequence, then

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

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

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

$$T_a = \left\{ t_i^I, \dots, t_{N_a}^V \right\}, \quad (3)$$

where t_i^I denotes the sub-template for the i th 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,

$$s(T_a, T_b) = \frac{\sum_{t_i \in T_a, t_j \in T_b} s(t_i, t_j) e^{\beta s(t_i, t_j)}}{\sum_{t_i \in T_a, t_j \in T_b} e^{\beta s(t_i, t_j)}}, \quad (4)$$

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° to -10° and $+10^\circ$ to $+90^\circ$ at a stride of 10° 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, respectively. 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° , 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 China Scholarship Council (CSC) grant 201503170248, and Shenzhen Qianhai DeepvisionX Technology Co.,Ltd, China.

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.

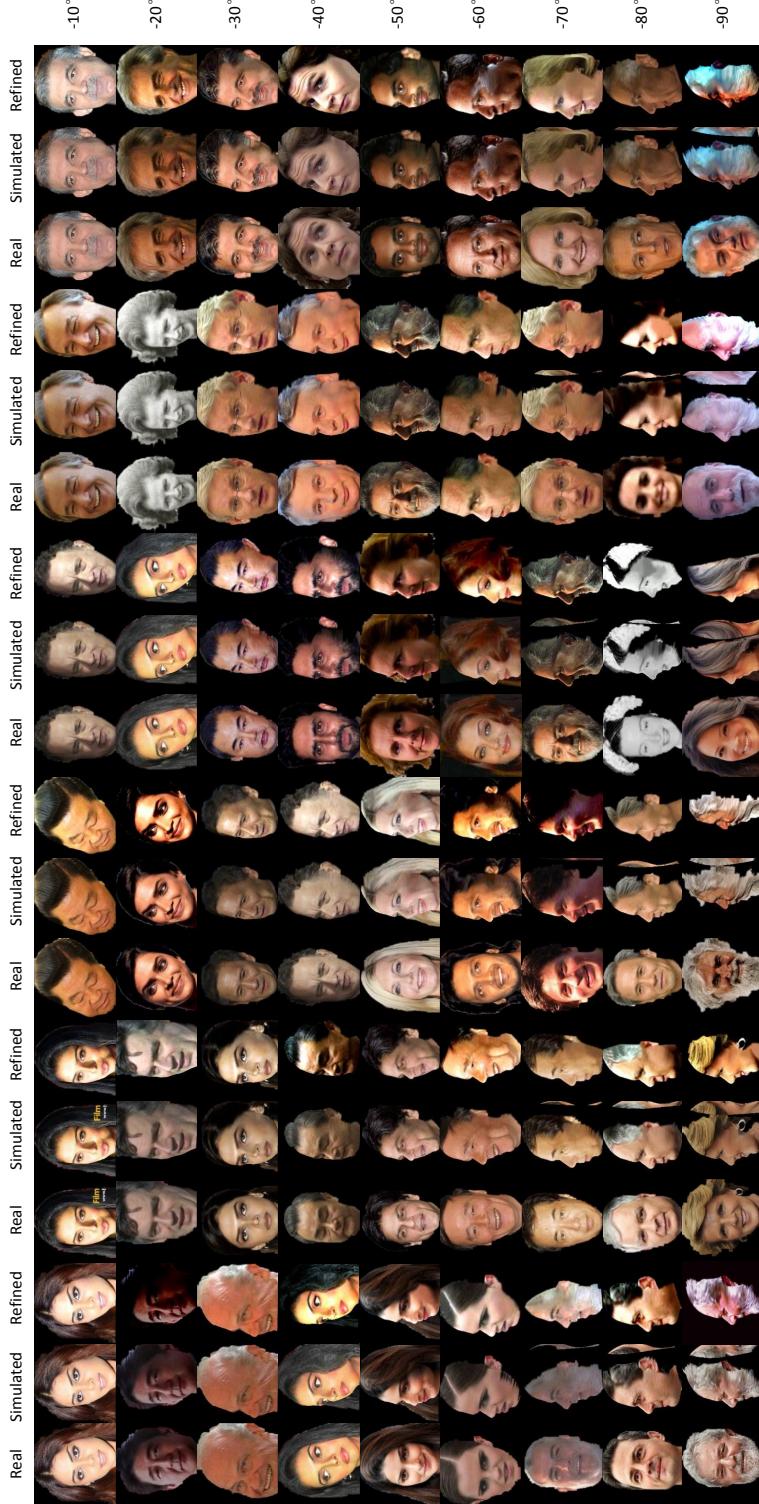


Figure 2: Refined results of DA-GAN under various poses with yaw angles ranging from -90° to -10° at a stride of 10° .

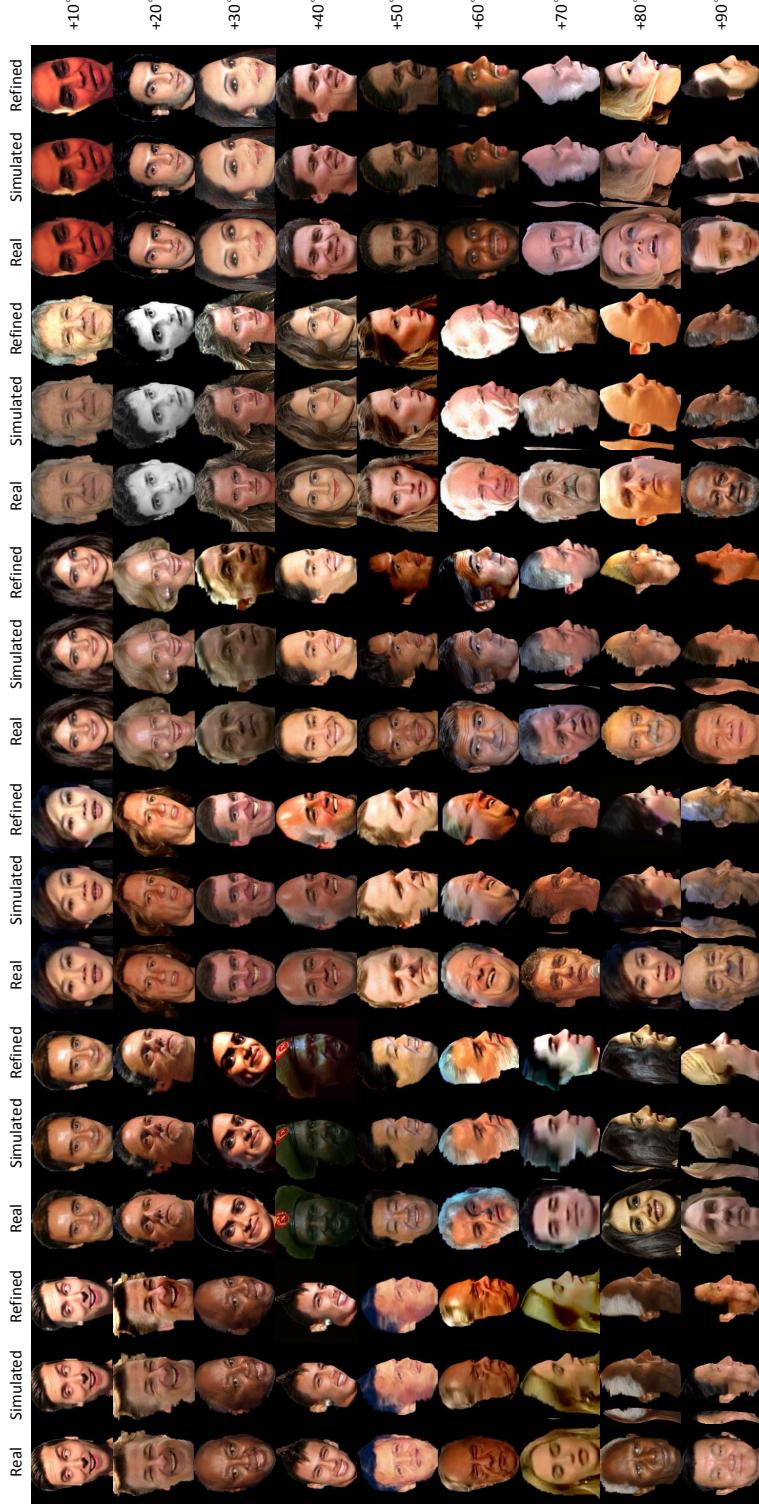


Figure 3: Refined results of DA-GAN under various poses with yaw angles ranging from $+10^\circ$ to $+90^\circ$ at a stride of 10° .

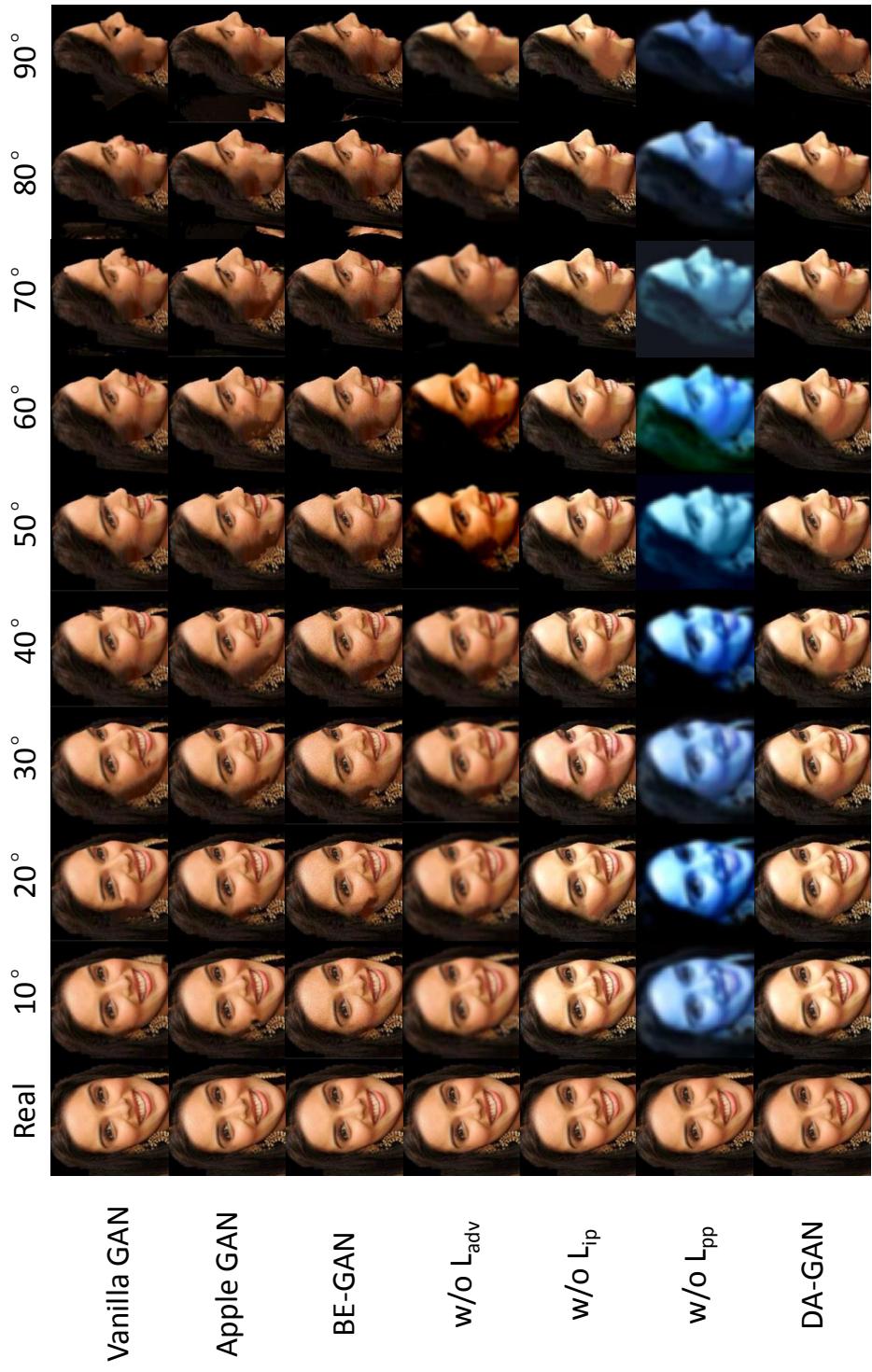


Figure 4: Qualitative results comparison of DA-GAN with state-of-the-art GANs and three different network settings.

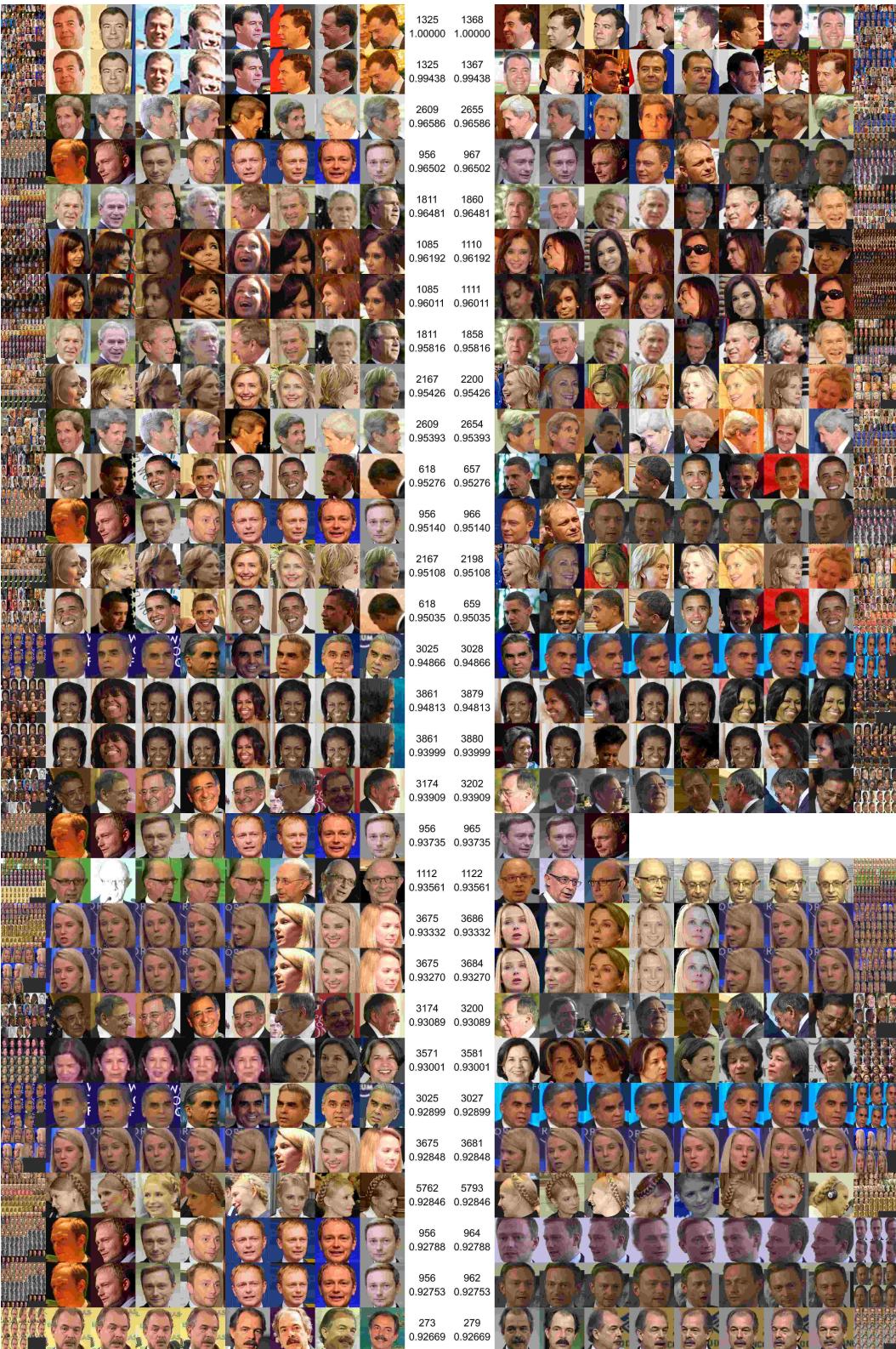


Figure 5: Verification results analysis for best matched cases on IJB-A split1.

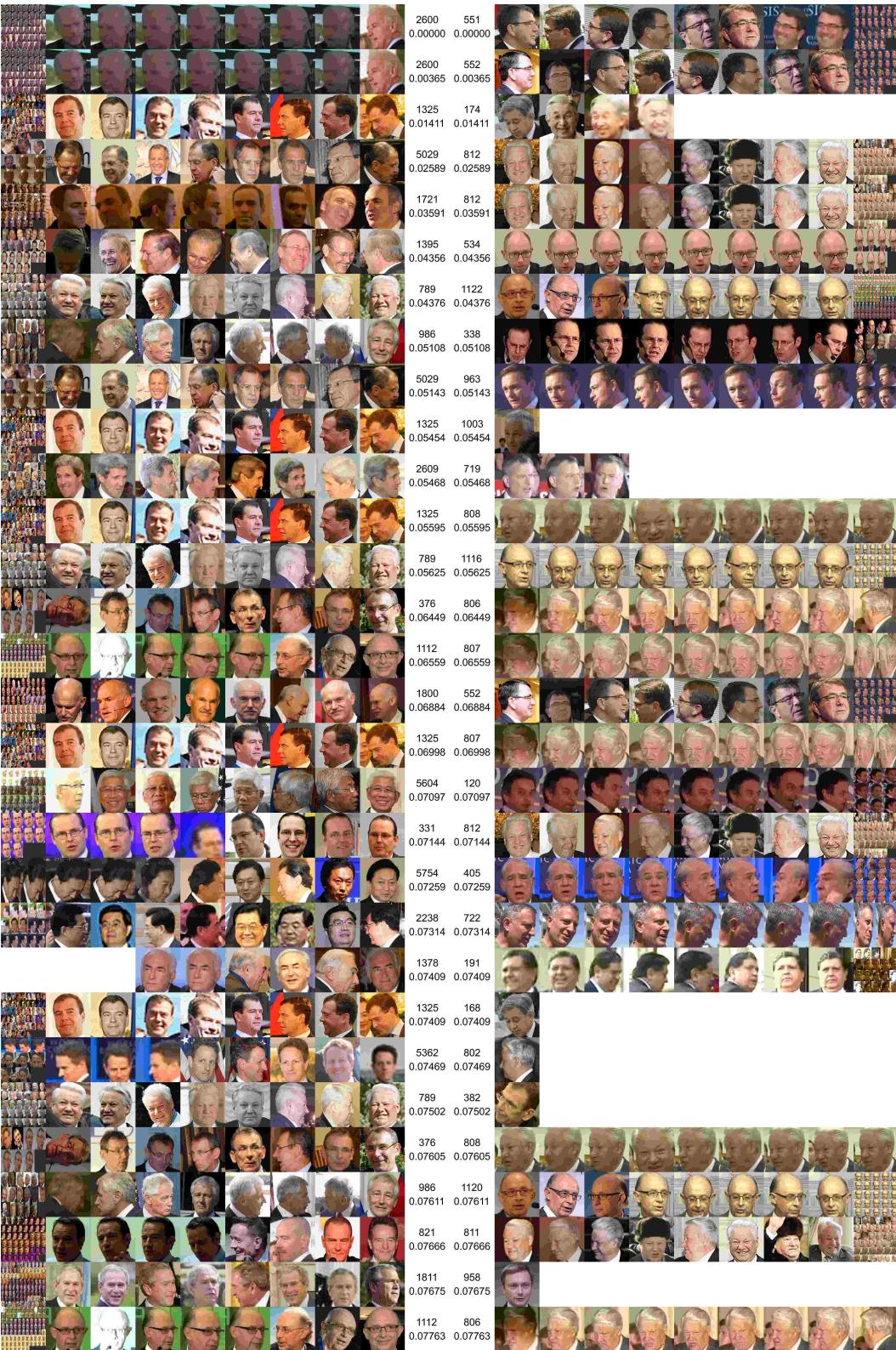


Figure 6: Verification results analysis for best non-matched cases on IJB-A split1.



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

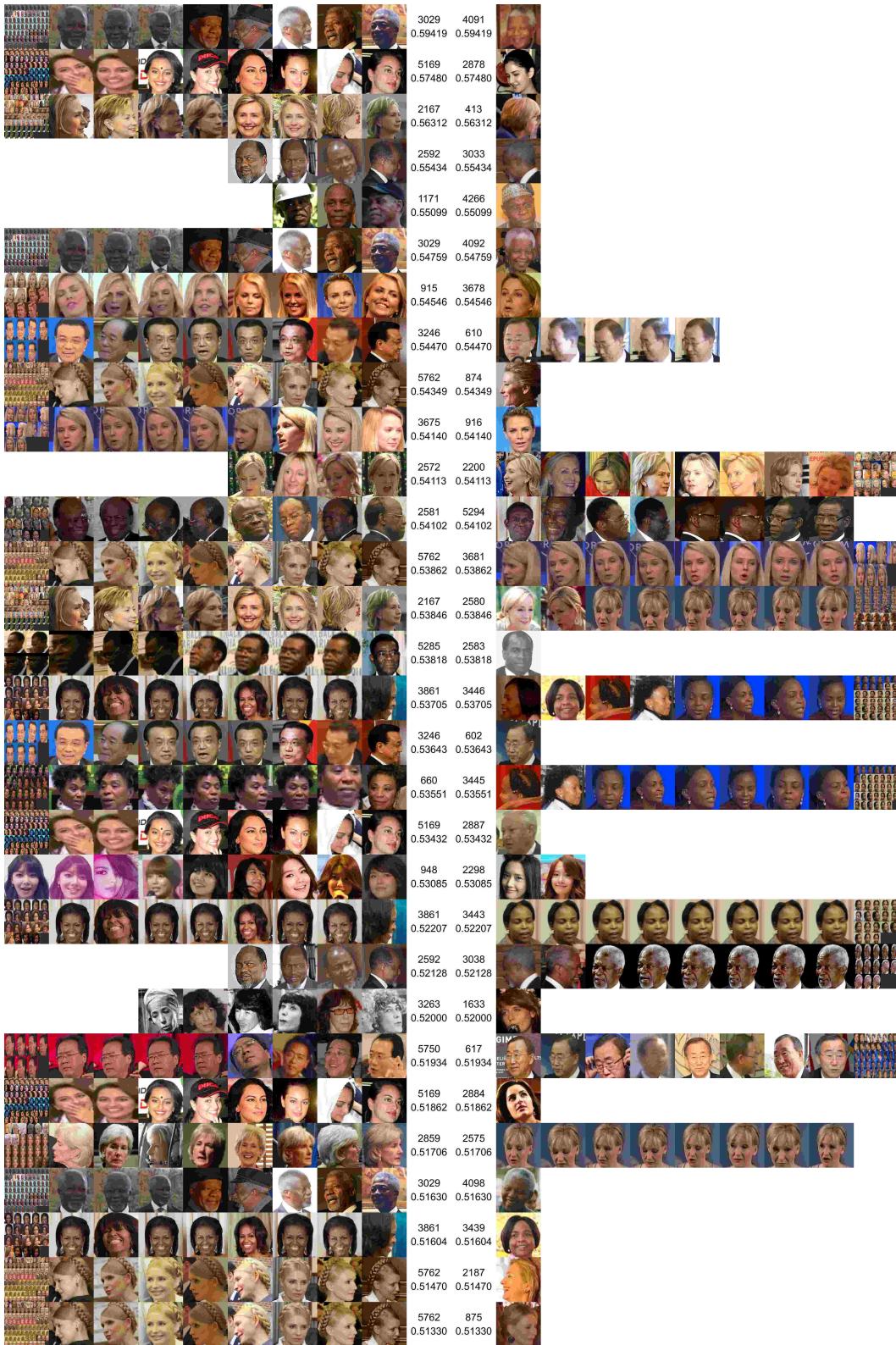


Figure 8: Verification results analysis for worst non-matched cases on IJB-A split1.

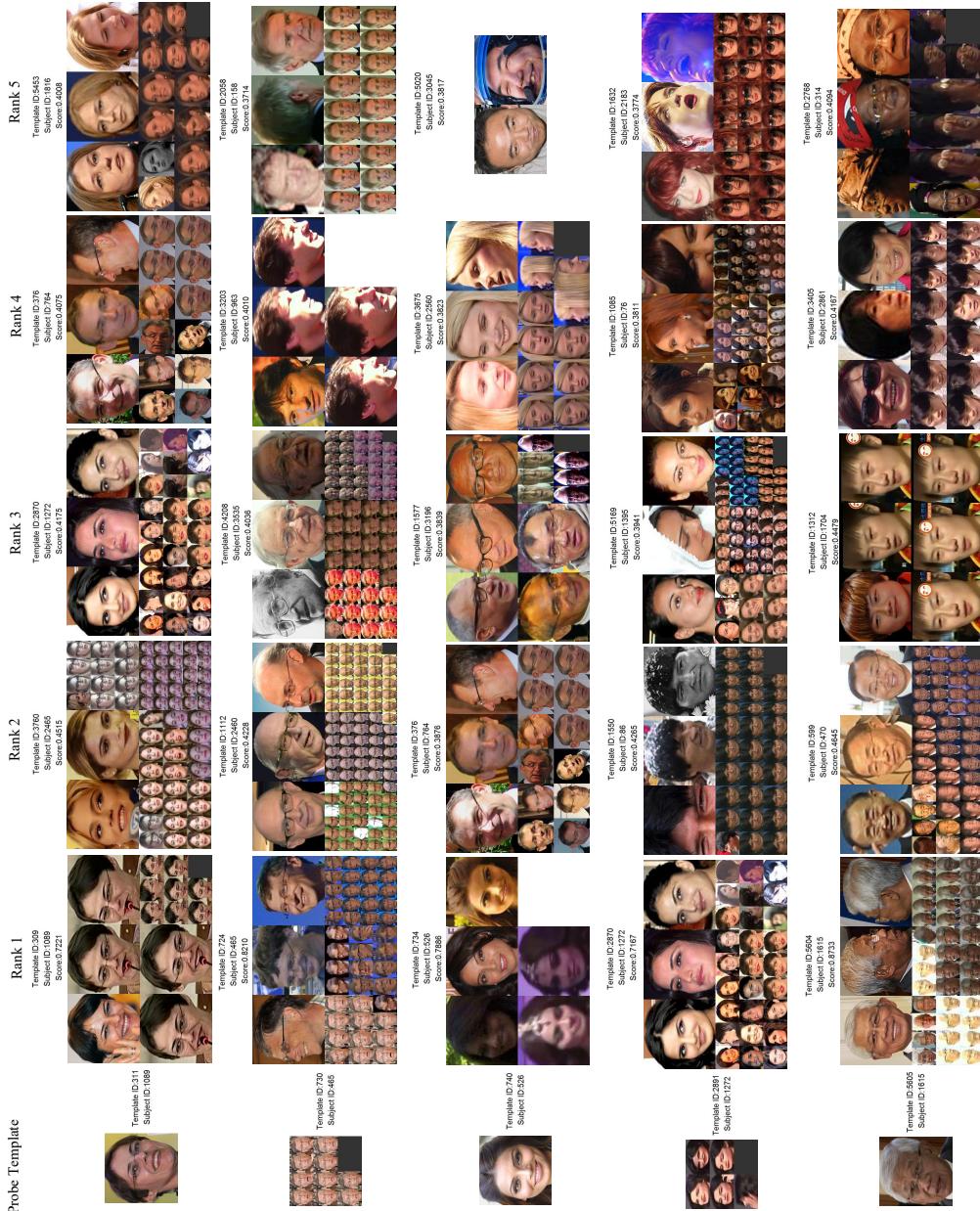


Figure 9: Identification results analysis on IJB-A split1.