Supplementary Material for Unsupervised Face Normalization with Extreme Pose and Expression in the Wild

Yichen Qian[1,2], Weihong Deng[1*], Jiani Hu[1]
[1]Beijing University of Posts and Telecommunications
[2]AI Labs, Didi Chuxing, Beijing 100193, China
{mx54039q, whdeng, jnhu}@bupt.edu.cn

Abstract

In this supplementary material, we present fully detailed information on 1) learning algorithm of the proposed Face Normalization Model (FNM); 2) details on IJB-A database [5]; 3) detailed network architectures; 4) training details; 5) more qualitative results of FNM.

1. Learning Algorithm of FNM

We summarize the detailed training procedures of our FNM in Algorithm. 1.

Algorithm 1 Learning algorithm of FNM

Input: Non-normal face images x, normal face images y, pretrained VGGFace2 network (G_{enc}), max number of epochs (ne), batch size (nb), number of network updates per epoch(ns), learning rate (lr), λ_1, λ_2, α, β_1, β_2, ϵ;
Output: the decoder of generator G_{dec} and discriminators D;
for e=1,...,ne **do**
  for s=1,...,ns **do**
    1. Optimize D;
    2. Optimize G_{dec};
    3. Visualize intermediate results
  end for
  Archive G_{dec} and D models for each training epoch;
end for

2. Details on IJB-A Database

IJB-A [5] 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 [5] 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 subjects 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, 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 [5].

3. Detailed Network Architectures

Architectures of our FNM are shown in Fig. 1. In the generator, we use transposed convolution layer [1] (kernel

*Corresponding author

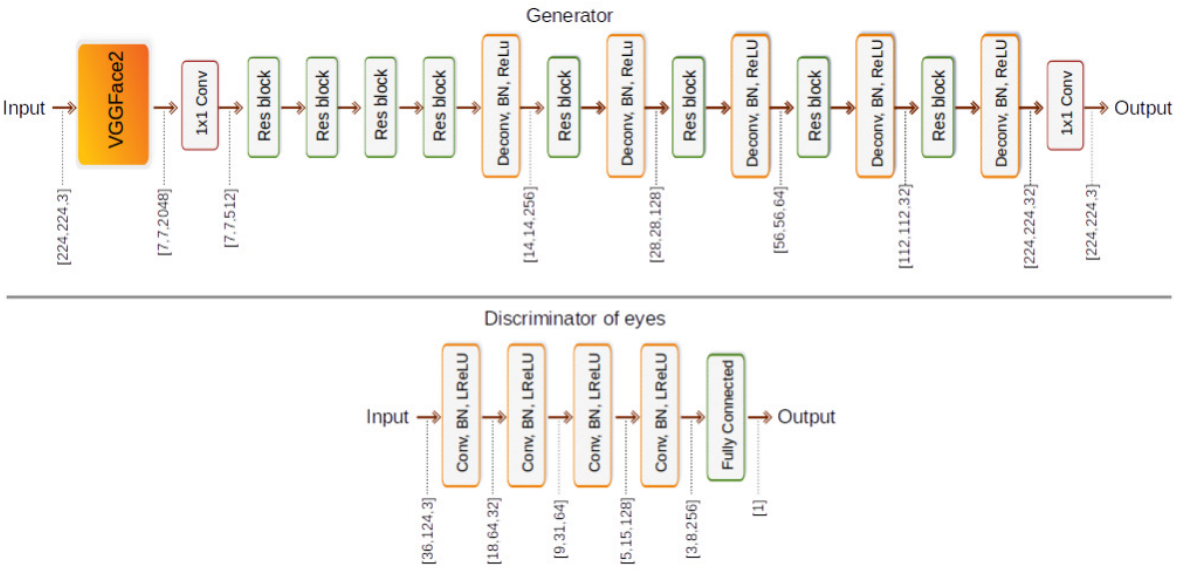

Figure 1. Architectures of Generator and Discriminator. In Generator, input is a colorful face image while output is a normalized face image with same identity. In Discriminator for eyes, input is a colorful image region of eyes while output is a logit. Architectures of other Discriminators for face, nose, mouth and image are similar. Dashed lines mark out positions where feature size changes.

= 4, stride = 2) to deconvolution feature, Rectified Linear Units (ReLU) to activate feature, Residual block [2] (Conv 3 $\times$ 3, BN [3], ReLU, Conv 3$\times$ 3, BN, skip connections, ReLU) to learn feature, 1 $\times$ 1 convolution to change feature channel. In the discriminators, we use convolution layer(kernel = 4, stride = 2) to downscale feature, Leaky Rectified Linear Units (LReLU) to activate feature, fully connected layer (FC) to transform feature. Batch Normalization is used throughout the generator and the discriminators.

We use the feature maps of "conv5_3" (7 $\times$ 7 $\times$ 2048) generated by VGGFace2, and then employ a 1 $\times$ 1 convolution and 4 Residual blocks. Each transposed convolution layer of G_{dec} is followed by one residual block [2]. Finally, we apply a 1 $\times$ 1 convolution to yield 224 $\times$ 224 $\times$ 3 RGB values. Architectures of Discriminator is similar in backbone but different in size of input region. The sizes of five attention regions (*i.e.*, image, face, eyes, nose, mouth) are (224,224), (112,124), (36,124), (66,44) and (30,74) respectively.

4. Training details

We train our FNM using Adam [4] with mini-batch; set the minibatch size to 16; W = 224, H = 224, C = 3; set learning rate to 10^{-4}; set the weight decay to 10^{-6}; set λ_1 = 10, λ_2 = 0.001; alternatively optimize discriminators D and generator G for each mini-batch.

5. Addition Results on IJB-A

We visualize the high-resolution face normalization results of FNM on unconstrained dataset IJB-A in Fig. 2 and Fig. 3. Our FNM presents a good identity preserving quality while producing photorealistic face normalization under large-pose, low resolution, occlusion and other complicated conditions. In addition, we take face images of the same person in different views across pose, lighting, expression and background in Fig. 3. The results of our FNM shows a highly consistency in preserving identity.

References

[1] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[5] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015. 1

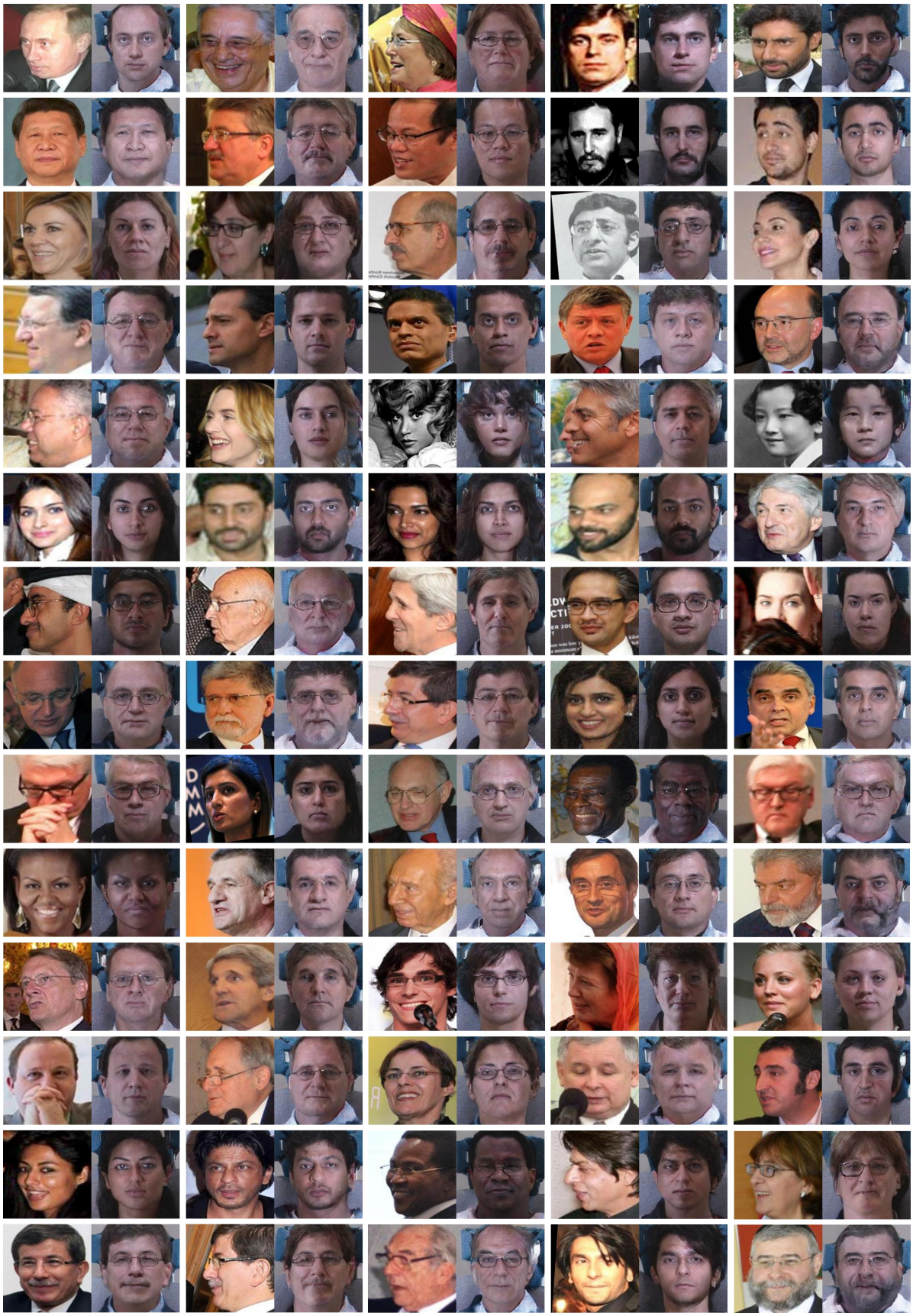

Figure 2. The high-resolution face normalization results of FNM on unconstrained dataset IJB-A.

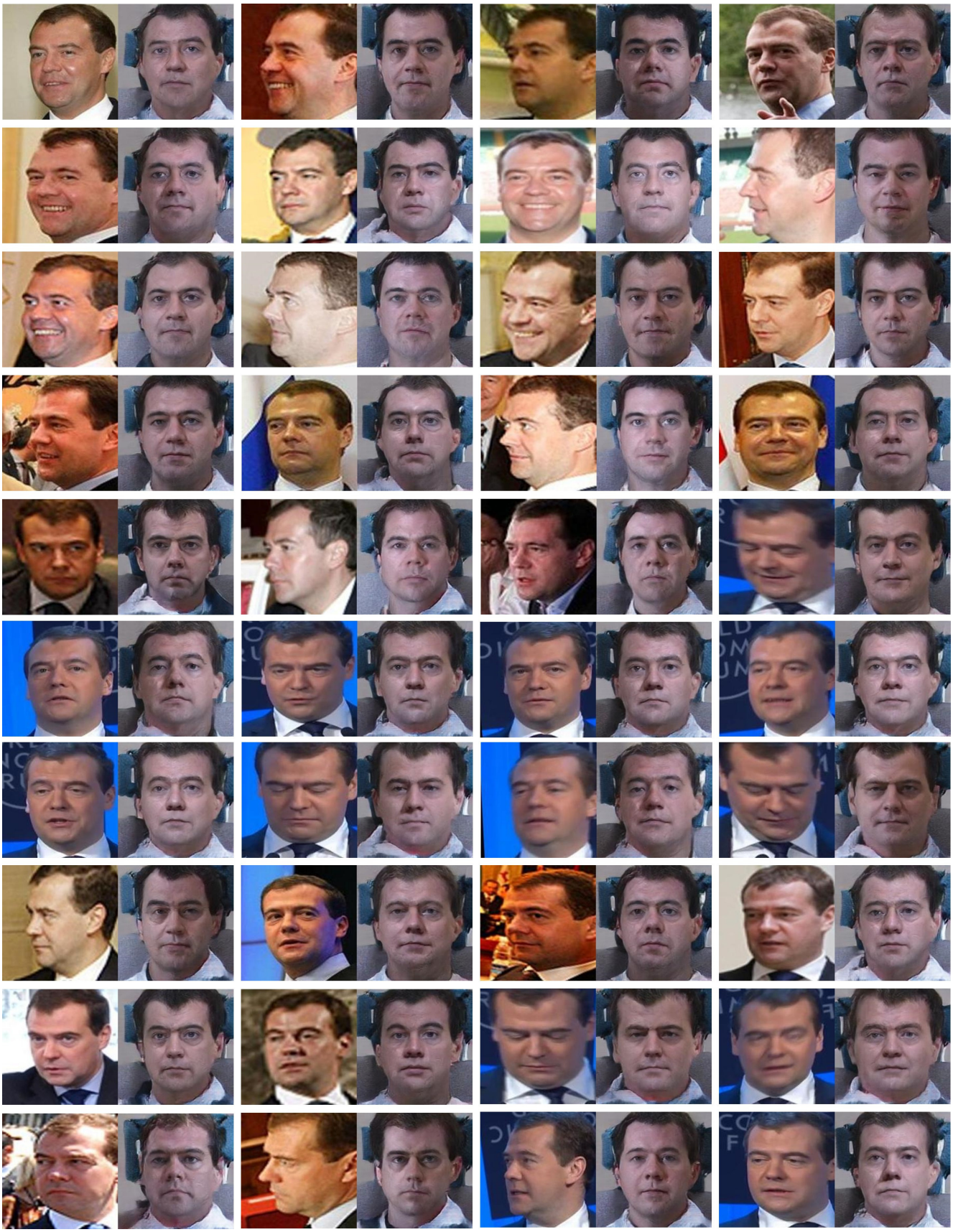

Figure 3. The face normalization results of FNM under various views of the same person.