Supplementary Materials for Pose-Guided Photorealistic Face Rotation

Yibo Hu[1,2], Xiang Wu[1], Bing Yu[3], Ran He[1,2*], Zhenan Sun[1,2]

[1]CRIPAC & NLPR & CEBSIT, CASIA [2]University of Chinese Academy of Sciences
[3]Noah's Ark Laboratory, Huawei Technologies Co., Ltd.
{yibo.hu, xiang.wu}@cripac.ia.ac.cn, yubing5@huawei.com, {rhe, znsun}@nlpr.ia.ac.cn

1. Network Architectures

The detailed architecture of pose-guided generator G_{θ_G}, including encoder and decoder, is shown in Table 1. In the encoder, each convolution layer is followed by one residual block [3] and the output of $fc1$ is operated by maxout [1]. The decoder contains three parts. The first part is a simple deconvolution structure to upsample the features $fc2$. The second part consists of deconvolution layers stacked for reconstruction and each of them is followed by two residual blocks. The third part involves some convolution layers for recovering different scales of face images.

For the couple-agent discriminator, the details are shown in Table 2, following the architectures in [5]. The $convk$ contains 4×4 convolution, instance normalization and leaky ReLU. For the last layer, we apply a sigmoid function to produce probabilistic output for optimization.

Table 1. Architecture of pose-guided generator G_{θ_G}.

Layer	Input	Filter/Stride	Output Size
conv0	I^a, P^a, P^b	$7 \times 7/1$	$128 \times 128 \times 64$
conv1	conv0	$5 \times 5/2$	$64 \times 64 \times 64$
conv2	conv1	$3 \times 3/2$	$32 \times 32 \times 128$
conv3	conv2	$3 \times 3/2$	$16 \times 16 \times 256$
conv4	conv3	$3 \times 3/2$	$8 \times 8 \times 512$
flatten, fc1	conv4	-	512
maxout	fc1	-	256
fc2, reshape	maxout	-	$8 \times 8 \times 64$
dc0_1	fc2	$4 \times 4/4$	$32 \times 32 \times 32$
dc0_2	dc0_1	$2 \times 2/2$	$64 \times 64 \times 16$
dc0_3	dc0_2	$2 \times 2/2$	$128 \times 128 \times 8$
dc1	fc2, conv4	$2 \times 2/2$	$16 \times 16 \times 512$
dc2	dc1, conv3	$2 \times 2/2$	$32 \times 32 \times 256$
dc3	dc2, conv2, I^a, dc0_1	$2 \times 2/2$	$64 \times 64 \times 128$
dc4	dc3, conv1, I^a, dc0_2	$2 \times 2/2$	$128 \times 128 \times 64$
conv5	dc2	$3 \times 3/1$	$32 \times 32 \times 3$
conv6	dc3	$3 \times 3/1$	$64 \times 64 \times 3$
conv7	dc4, conv0, I^a, dc0_3	$5 \times 5/1$	$128 \times 128 \times 64$
conv8	conv7	$3 \times 3/1$	$128 \times 128 \times 32$
conv9	conv8	$3 \times 3/1$	$128 \times 128 \times 3$

Table 2. Architecture of the discriminators $D_{\theta_{ii}}$ and $D_{\theta_{pe}}$.

Layer	Input	Filter/Stride	Output Size
conv0	$\hat{I}^b$, $I^a/\hat{I}^b$, P^b	$4 \times 4/2$	$64 \times 64 \times 64$
conv1	conv0	$4 \times 4/2$	$32 \times 32 \times 128$
conv2	conv1	$4 \times 4/2$	$16 \times 16 \times 256$
conv3	conv2	$4 \times 4/2$	$8 \times 8 \times 512$
conv4	conv3	$4 \times 4/1$	$7 \times 7 \times 512$
conv5	conv4	$4 \times 4/1$	$6 \times 6 \times 1$

2. Additional Results on Multi-PIE

Additional synthesis results on the Multi-PIE dataset are presented in Fig. 2, Fig. 3, Fig. 4 and Fig. 5. Note that all the synthesized images are 128×128. Fig. 2 and Fig. 3 demonstrate that CAPG-GAN is robust to illumination changes. Despite extreme illumination variations, the skin tone, global structure and facial details are consistent across different illuminations. However, we find that the facial contours are hardly recovered when frontalization. It may be because rotating a profile face to the frontal view is an ill-posed problem and a profile face only contains part of the information of the frontal face. As shown in Fig. 4, our CAPG-GAN has the ability to rotate a face to an arbitrary pose, even though it is trained only with frontal and profile face pairs. The results in Fig. 4 suggest its potential usage for face data augmentation. Usually, there is illumination shift when faces are rotated from small poses to large poses. Fig. 5 further gives the evidence of pose guided characteristic, where we interpolate the positions of the target landmarks, guiding to rotate the source faces to the target poses. We can find the global and local information is consistent across different poses, despite that the target poses are not seen during training.

3. Additional Results on LFW

Additional synthesis results and visual comparison with other methods on the LFW dataset are shown in Fig. 1. Note that CAPG-GAN is only trained on Multi-PIE and directly tested on LFW. The face rotation task is more challeng-

*corresponding author

Figure 1. Synthesis results of CAPG-GAN and its competitors on the LFW dataset. (a) Source images. (b) CAPG-GAN. (c) TP-GAN [4]. (d) HPEN [6]. (e) LFW-3D [2].

ing on LFW than Multi-PIE because there are various occlusions in LFW incurred by sunglasses, fingers and etc. It means that one should deal with both face rotation and completion during frontalization. Although CAPG-GAN is merely trained on Multi-PIE and all the occlusions in LFW are unseen during training, it can still tackle these occlusions well as shown in Fig. 1. CAPG-GAN obtains better visual results than [6, 2]. It can preserve identity information well for large poses albeit some small local details are not reconstructed very well. In the future, we will consider to use large-scale in-the-wild training data with various occlusions to train our method and utilize 3D models to further improve the synthesis quality.

References

[1] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. In *ICML*, 2013.

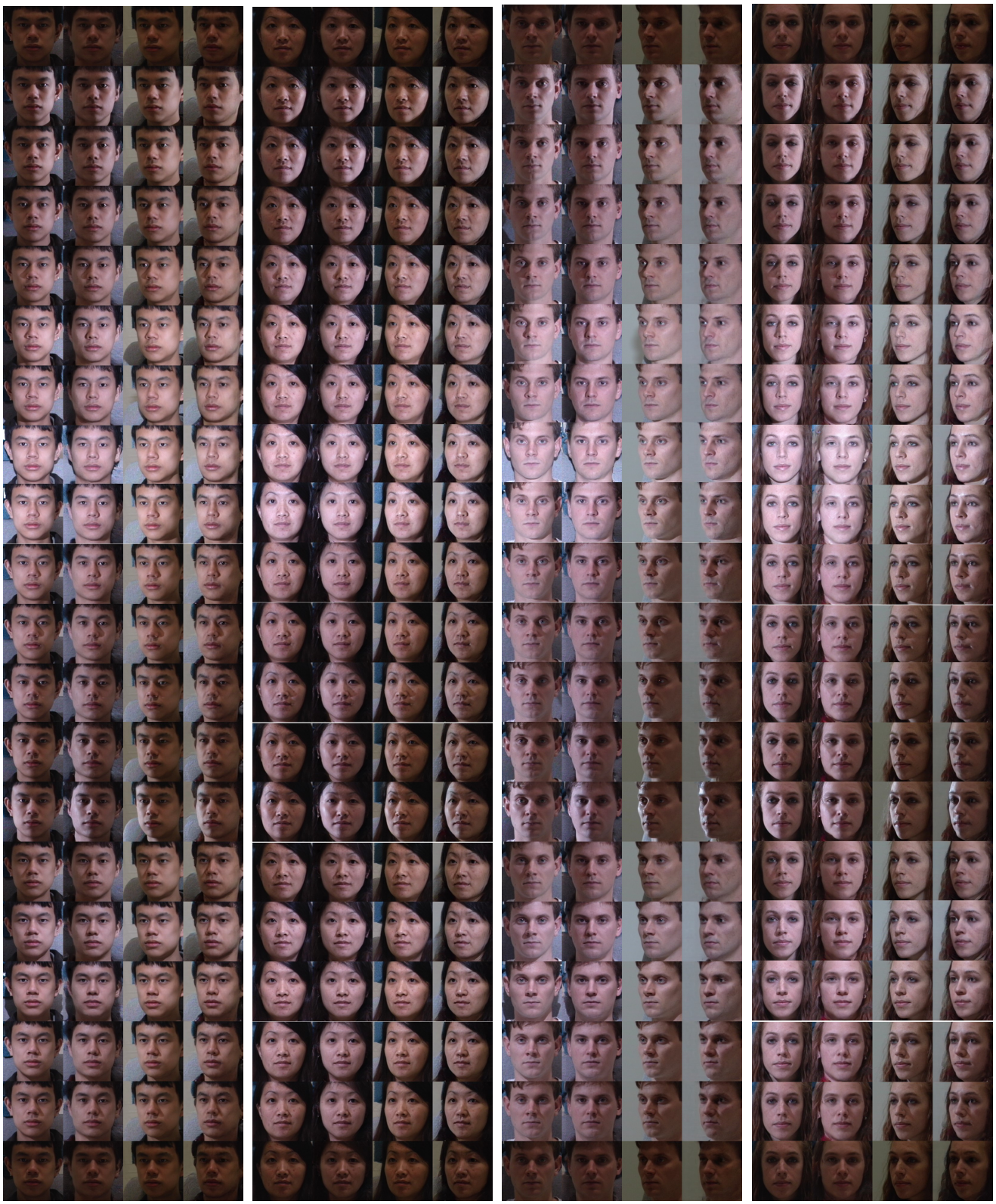
[2] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017.

[5] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[6] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015.

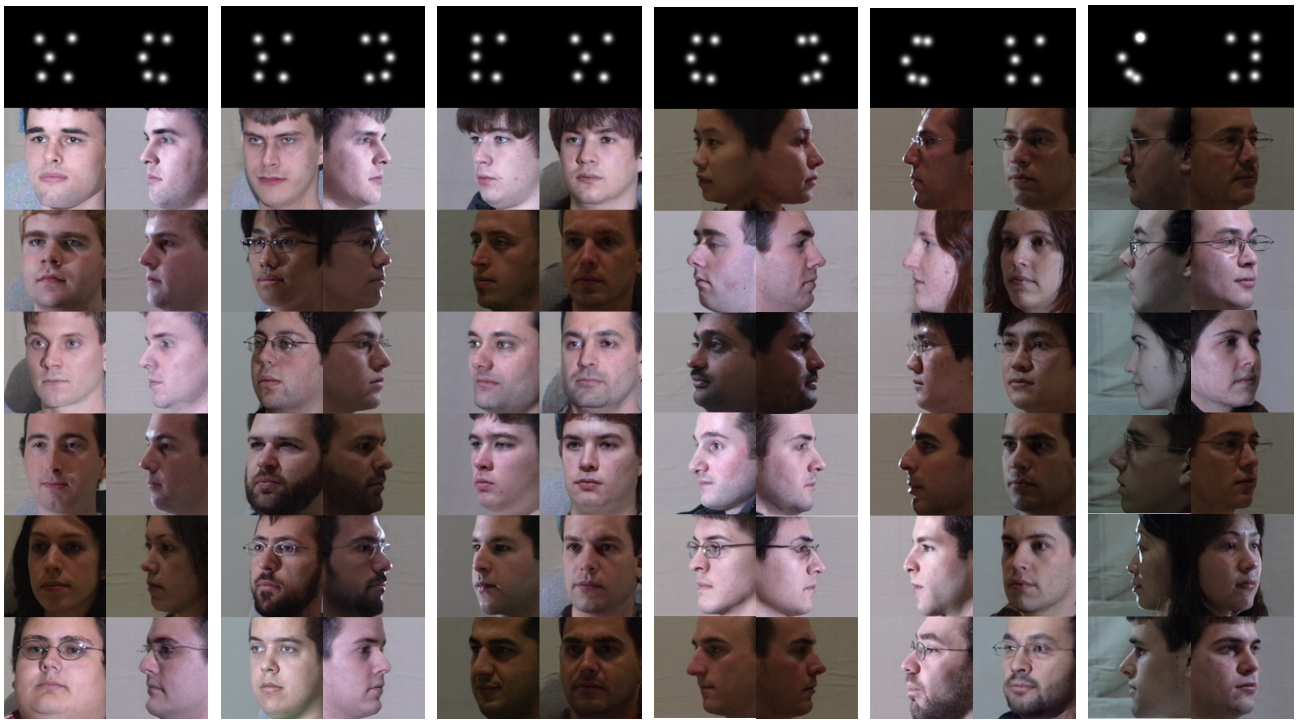


(a) (b) (c) (d)

Figure 2. Synthesis results under various illuminations (small poses). Each row shares the same illumination and each column shows the images of the same subject under the same pose. The first and third images in each column are ground truth, and the second and forth images are synthesized results.

(a) (b) (c) (d)

Figure 3. Synthesis results under various illuminations (large poses). Each row shares the same illumination and each column shows the images of the same subject under the same pose. The first and third images in each column are ground truth, and the second and forth images are synthesized results.



(a) $+15° \rightarrow +60°$ (b) $+30° \rightarrow -60°$ (c) $+45° \rightarrow +15°$ (d) $+60° \rightarrow -75°$ (e) $+75° \rightarrow +30°$ (f) $+90° \rightarrow -45°$

Figure 4. Synthesis results of various poses. For a pair of face images in one column, the left is the source image and the right is the synthesized one guided by the landmarks above.

(a) Source (b) $7.5°$ (c) $15°$ (d) $22.5°$ (e) $30°$ (f) $37.5°$ (g) $45°$ (h) $52.5°$ (i) $60°$ (j) $67.5°$ (k) $75°$ (l) $82.5°$

Figure 5. Synthesis results of landmark interpolation.