Supplementary Materials for M^{2}FPA: A Multi-Yaw Multi-Pitch High-Quality Dataset and Benchmark for Facial Pose Analysis

Peipei Li[1,2], Xiang Wu[1], Yibo Hu[1], Ran He[1,2]*, Zhenan Sun[1,2]

[1]CRIPAC & NLPR & CEBSIT, CASIA [2]University of Chinese Academy of Sciences
Email: {peipei.li, yibo.hu}@cripac.ia.ac.cn, alfredxiangwu@gmail.com, {rhe, znsun}@nlpr.ia.ac.cn

In this supplementary material, we first introduce the network architectures of the generator and discriminators in our method. Then we present the ablation study in Section 2. Additional in-the-wild experiments on LFW and CelebA-HQ are shown in Section 3 and 4, respectively. 256×256 frontalization results for all the 57 poses are given in Section 5. Furthermore, in Section 6, we conduct face frontalization of 512×512 resolution on the new M^{2}FPA database, which reveals the superiority of M^{2}FPA.

1. Network Architecture

Our generator G_{θ_G} adopts an encoder-decoder architecture. Taking 256×256 resolution as an example, the detailed structure of G_{θ_G} is listed in Table 1. In the encoder, each convolution layer is followed by one residual block. In the decoder, there are three parts. The first is a simple deconvolution structure to upsample the fc2 features. The second part contains stacked deconvolution layers for reconstruction and each of them is followed by two residual blocks. The third one involves some convolution layers for recovering different scales of face images.

The detailed structures of the global discriminator $D_{\theta_{D1}}$ and the parsing guided local discriminator $D_{\theta_{D2}}$ are shown in Tables 2 and 3, respectively. Each $convk$ in $D_{\theta_{D1}}$ and $D_{\theta_{D2}}$ contains a 3×3 convolution layer, an instance normalization layer and a leaky ReLU layer. The last layers in $D_{\theta_{D1}}$ and $D_{\theta_{D2}}$ produce probabilistic outputs by sigmoid functions.

Note that, we also employ the same network architectures for experiments of 128×128 resolution (in the main text) and 512×512 resolution (in this supplementary material), except for the channel numbers of $fc1$ and $fc2$.

Table 1. Structure of the generator G_{θ_G}.

Layer	Input	Filter Size	Output Size
$conv0$	X	$7 \times 7/1$	$256 \times 256 \times 64$
$conv1$	$conv0$	$5 \times 5/2$	$128 \times 128 \times 64$
$conv2$	$conv1$	$3 \times 3/2$	$64 \times 64 \times 128$
$conv3$	$conv2$	$3 \times 3/2$	$32 \times 32 \times 256$
$conv4$	$conv3$	$3 \times 3/2$	$16 \times 16 \times 512$
$fc1$	$conv4$	-	512
$maxout$	$fc1$	-	256
$fc2$	$maxout$	-	$16 \times 16 \times 64$
$dec0\_1$	$fc2$	$4 \times 4/4$	$64 \times 64 \times 32$
$dec0\_2$	$dec0\_1$	$2 \times 2/2$	$128 \times 128 \times 16$
$dec0\_3$	$dec0\_2$	$2 \times 2/2$	$256 \times 256 \times 8$
$dec1$	$fc2, conv4$	$2 \times 2/2$	$32 \times 32 \times 512$
$dec2$	$dec1, conv3$	$2 \times 2/2$	$64 \times 64 \times 256$
$dec3$	$dec2, conv2, X, dec0\_1$	$2 \times 2/2$	$128 \times 128 \times 128$
$dec4$	$dec3, conv1, X, dec0\_2$	$2 \times 2/2$	$256 \times 256 \times 64$
$conv5$	$dec2$	$3 \times 3/1$	$64 \times 64 \times 64$
$conv6$	$dec3$	$3 \times 3/1$	$128 \times 128 \times 32$
$conv7$	$dec4, conv0, X, dec0\_3$	$5 \times 5/1$	$256 \times 256 \times 3$
$conv8$	$conv7$	$3 \times 3/1$	$256 \times 256 \times 3$
$conv9$	$conv8$	$3 \times 3/1$	$256 \times 256 \times 3$

Table 2. Structure of the discriminator $D_{\theta_{D1}}$.

Layer	Input	Filter Size	Output Size
$conv1$	$Y/\hat{Y}$	$3 \times 3/2$	$128 \times 128 \times 64$
$conv2$	$conv1$	$3 \times 3/2$	$64 \times 64 \times 128$
$conv3$	$conv2$	$3 \times 3/2$	$32 \times 32 \times 256$
$conv4$	$conv3$	$3 \times 3/2$	$16 \times 16 \times 512$
$conv5$	$conv4$	$3 \times 3/2$	$8 \times 8 \times 512$
$conv6$	$conv5$	$3 \times 3/2$	$4 \times 4 \times 512$
$conv7$	$conv6$	$1 \times 1/1$	$4 \times 4 \times 1$

2. Ablation Study

In this section, we report both qualitative visualization results and quantitative recognition results for a comprehensive comparison as the ablation study. Figure 1 presents visual comparisons between our method and its four incomplete variants on the new M^{2}FPA database. Without the $L_{adv_{1,2}}$ loss, the synthesized faces are obviously blur. Without the L_{ip} loss, much identity information is lost during face frontalization. Without L_{tv} loss, there are more artifacts on the synthesized faces. Specially, without the L_{adv2} loss, we observe that the structures of facial features are quite different from the ground truth, where the eyes and mouth have deformations. These indicate that the parsing

*corresponding author

Table 3. Structure of the discriminator $D_{\theta_{D2}}$.

Layer	Input	Filter Size	Output Size
h_conv1	$Y_h/\hat{Y}_h$	3 × 3/2	128 × 128 × 64
h_conv2	h_conv1	3 × 3/2	64 × 64 × 128
h_conv3	h_conv2	3 × 3/2	32 × 32 × 256
h_conv4	h_conv3	3 × 3/2	16 × 16 × 512
h_conv5	h_conv4	3 × 3/2	8 × 8 × 512
s_conv1	$Y_s/\hat{Y}_s$	3 × 3/2	128 × 128 × 64
s_conv2	s_conv1	3 × 3/2	64 × 64 × 128
s_conv3	s_conv2	3 × 3/2	32 × 32 × 256
s_conv4	s_conv3	3 × 3/2	16 × 16 × 512
s_conv5	s_conv4	3 × 3/2	8 × 8 × 512
f_conv1	$Y_f/\hat{Y}_f$	3 × 3/2	128 × 128 × 64
f_conv2	f_conv1	3 × 3/2	64 × 64 × 128
f_conv3	f_conv2	3 × 3/2	32 × 32 × 256
f_conv4	f_conv3	3 × 3/2	16 × 16 × 512
f_conv5	f_conv4	3 × 3/2	8 × 8 × 512
F_conv1	h, s, f_conv5	3 × 3/1	8 × 8 × 512
F_conv2	F_conv1	3 × 3/2	4 × 4 × 512
F_conv3	F_conv2	1 × 1/1	4 × 4 × 1

Table 4. Model comparisons: Rank-1 recognition rates (%) on M^{2}FPA.

Method	±15°	±30°	±45°	±60°	±75°	±90°
LightCNN-29 v2						
w/o $L_{adv1,2}$	99.8	99.7	99.4	97.3	86.1	63.1
w/o L_{tv}	99.8	99.6	99.5	97.9	88.6	67.1
w/o L_{ip}	99.9	99.7	99.0	96.9	86.3	56.5
w/o L_{adv2}	100	100	99.7	98.4	89.3	63.5
Ours	**100**	**100**	**99.9**	**98.4**	**90.6**	**67.6**
IR-50						
w/o $L_{adv1,2}$	99.7	99.3	98.3	94.9	82.1	44.9
w/o L_{tv}	99.4	99.4	98.5	96.2	87.7	52.0
w/o L_{ip}	99.2	99.0	98.3	95.3	83.8	43.4
w/o L_{adv2}	99.7	99.3	98.3	95.7	82.4	45.9
Ours	**99.5**	**99.5**	**99.0**	**97.3**	**89.6**	**55.8**

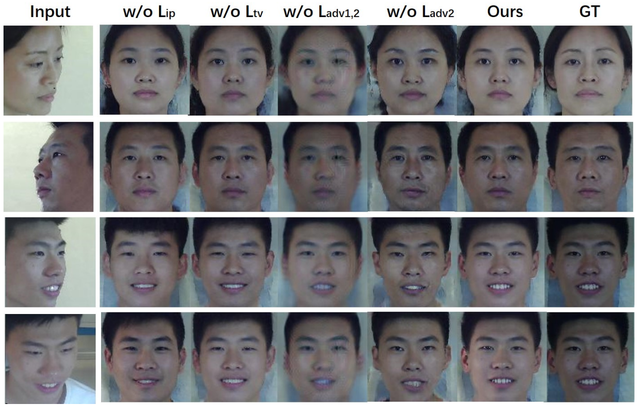

Figure 1. Model comparisons: synthesis results of our method and its variants.

guided local discriminator can ensure the local consistency between real and synthesized frontal images.

Table 4 further presents the Rank-1 performance of different variants of our method on M^{2}FPA. Similar to the visualization ablation study, we observe that the Rank-1 accuracy will decrease if one loss is removed. These phenomena indicate that each component in our method is essential for synthesizing photo-realistic frontal images.

Table 5. Face verification accuracy (ACC) and area-under-curve (AUC) results on LFW.

Method	ACC(%)	AUC(%)
Ferrari *et al.* [2]	-	94.29
LFW-3D[3]	93.62	88.36
LFW-HPEN[8]	96.25	99.39
FF-GAN[6]	96.42	99.45
CAPG-GAN[4]	99.37	99.90
Ours	**99.41**	**99.92**

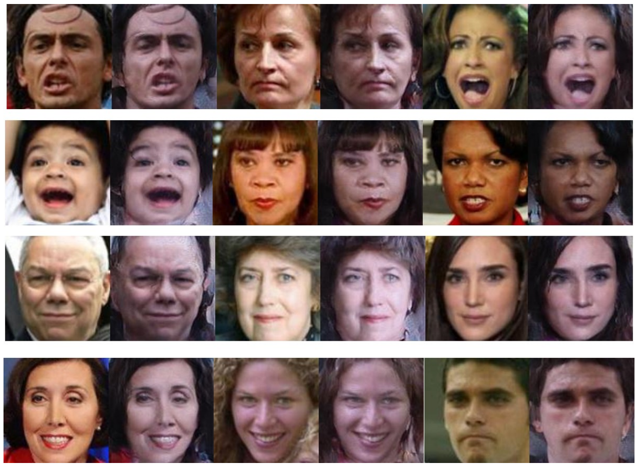

Figure 2. Visualization results on LFW. For each subject, the left is the input and the right is the frontalized result.

3. Additional Results on LFW

Additional frontalization results and comparisons with the previous methods on LFW are shown in Figure 2 and Figure 3, respectively. Same as TP-GAN [5] and CAPG-GAN [4], our model is only trained on Multi-PIE and tested on LFW. In Figure 2, for each subject, the input image is on the left and the frontalized result is on the right. We can observe that both the visual realism and the identity information are well preserved during frontalization. In addition, as shown in Figure 3, our method obtains good visualization results that are comparable to or better than the previous methods, including LFW-3D [3], LFW-HPEN [8], TP-GAN [5] and CAPG-GAN [4]. The quantitative results on LFW are presented in Table 5.

4. Additional Results on CelebA-HQ

CelebA-HQ [6] is a newly proposed high-quality database with small pose variations for face synthesis. We conduct additional experiments on CelebA-HQ to demonstrate the effectiveness of our method under such in-the-wild settings. We observe that the images in CelebA-HQ are almost frontal view. In order to take advantage of the high-quality images, following [1], we utilize a 3DMM model [7] to produce the paired profile images for each frontal image. We random choose 3,451 images as the testing set and the frontalization results of our method are presented in Fig-

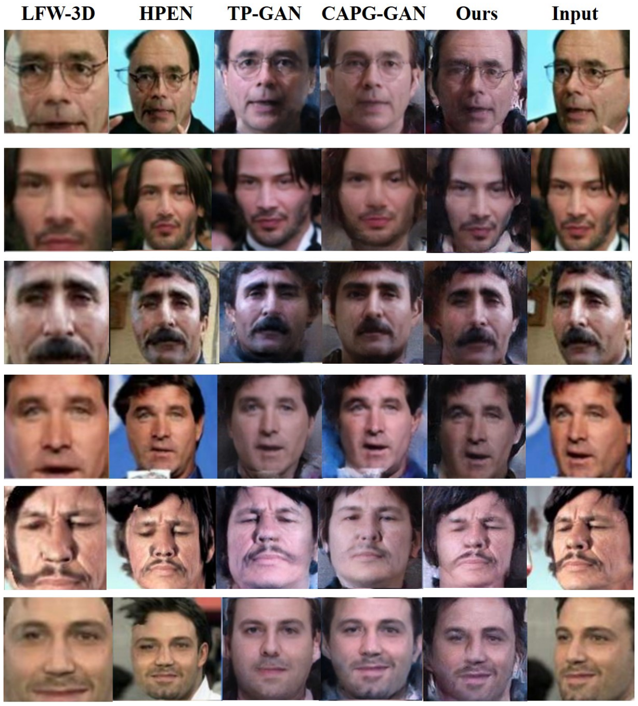

Figure 3. Visualization comparisons on LFW. For each subject, from left to right is the synthesized result of LFW-3D [3], HPEN [8], TP-GAN [5], CAPG-GAN [4], our method and the input image.

ure 4. Note that there are no overlap subjects between the training and testing sets.

5. Additional 256×256 Results on M^{2}FPA

Additional 256×256 frontalization results under 57 poses on M^{2}FPA are shown in Figure 5. For each subject, the top is the input with different poses and the bottom is the synthesized result. As expected, our method can frontalize the faces with sunglasses. In addition, we also observe that most frontalization results preserve the visual realism and the identity information well, even under extreme yaw and pitch poses.

6. Additional 512×512 Results on M^{2}FPA

Generating high-resolution results is significant to enlarge the application field of face rotation. However, the current facial pose analysis databases, which are collected in the constrained environment, only provide 128×128 images. Our proposed M^{2}FPA supports higher resolution up to 512×512 and contains various yaw and pitches angels. Additional 512×512 frontalization results of our method on M^{2}FPA are shown in Figure 6. We observe that our high resolution results have richer textures and look more plausible. We believe that the high-resolution M^{2}FPA can push forward the advance of facial pose analysis in mobile or surveillance applications.

References

[1] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In *NeurIPS*, 2018.

[2] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Effective 3d based frontalization for unconstrained face recognition. In *ICPR*, 2016.

[3] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

[4] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *CVPR*, 2018.

[5] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017.

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[7] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009.

[8] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015.

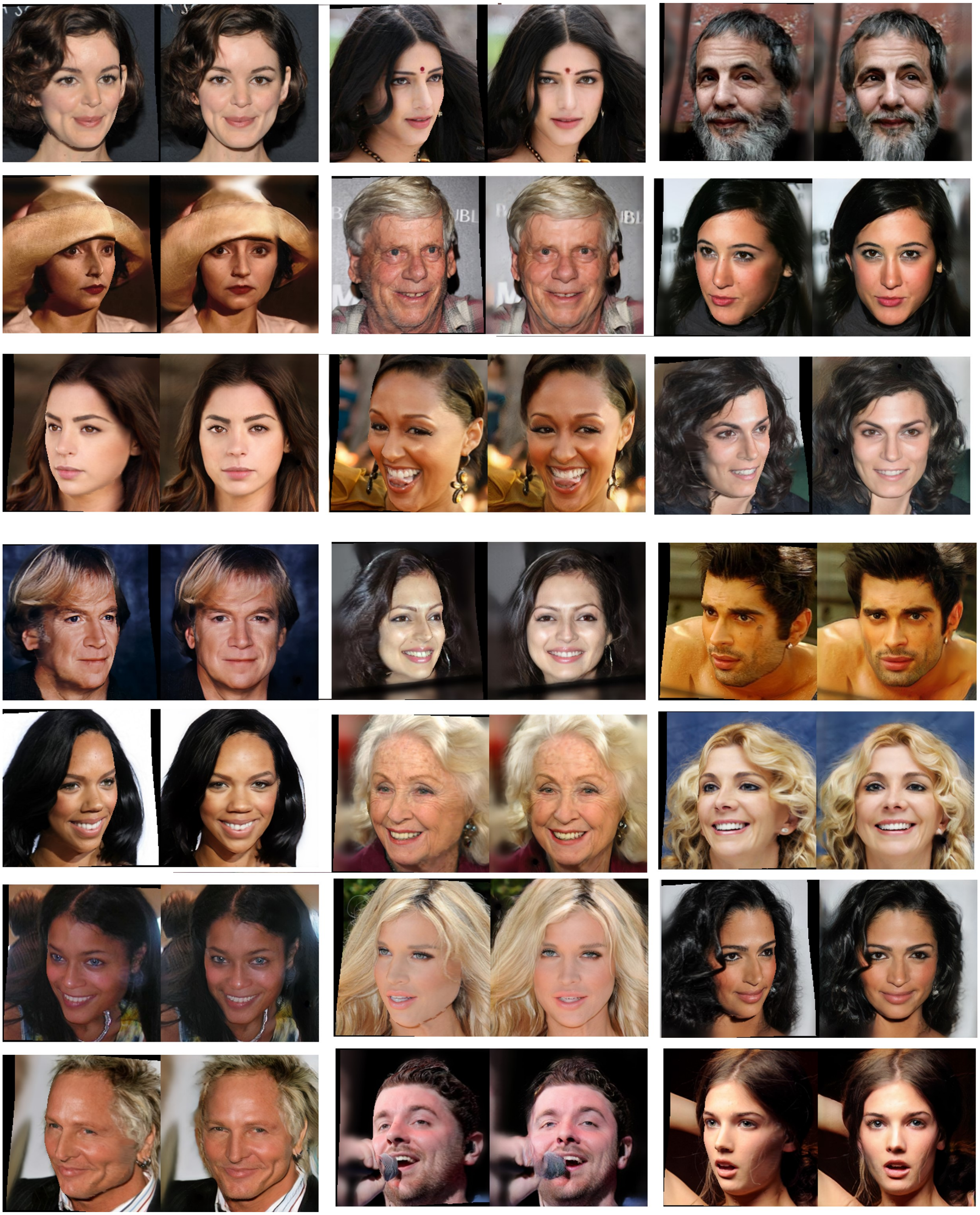

Figure 4. High-quality frontalization results on CelebA-HQ. For each subject, the left is the input and the right is the synthesized result.

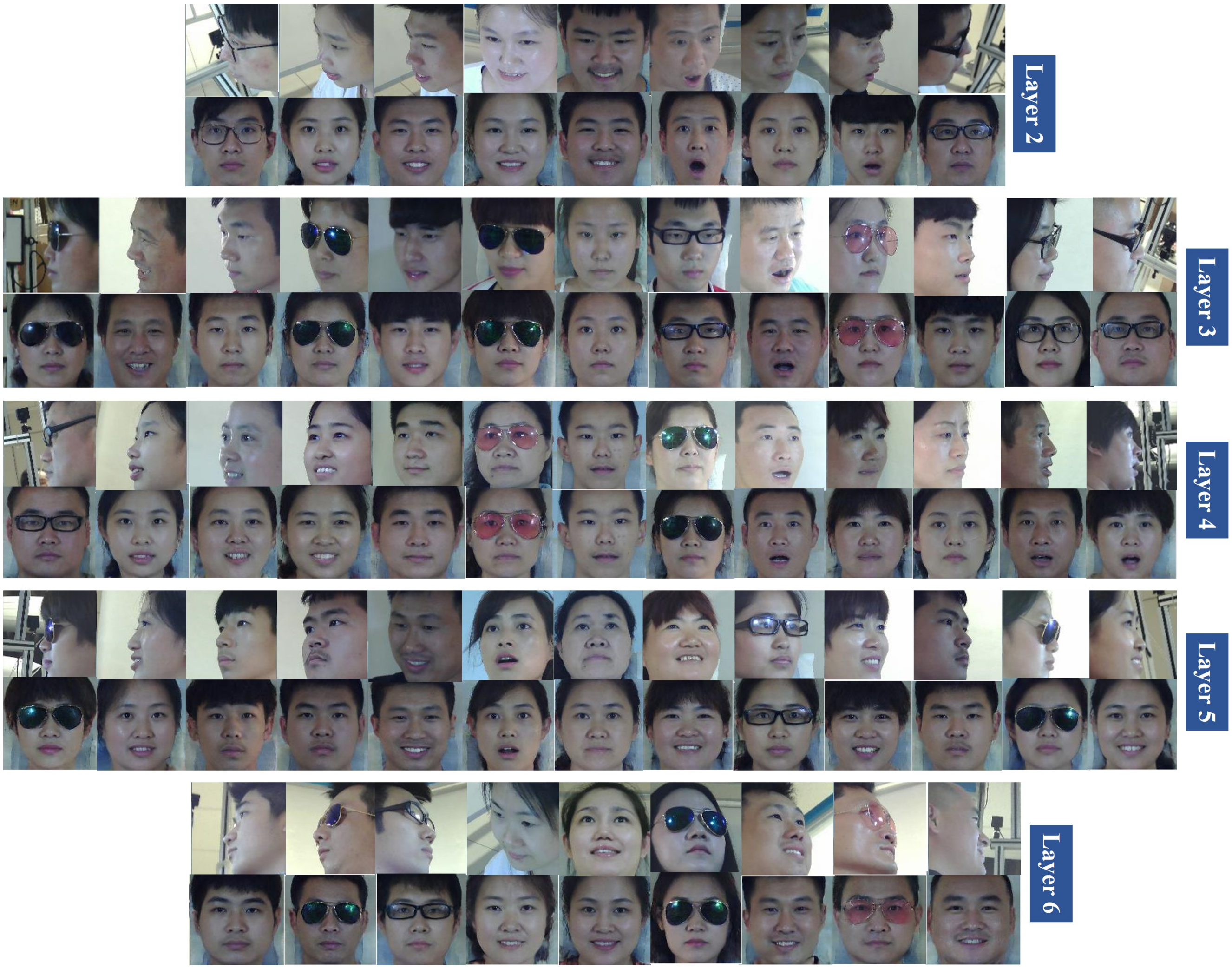

Figure 5. The 256×256 frontalization results of our method under 57 poses on M^{2}FPA. From top to bottom, the pitch angles of the Layer 2-6 are $+30^\circ$, $+15^\circ$, 0°, -15° and -30°, respectively. From left to right, the yaw angles are from -90° to $+90^\circ$. For each subject, the top is the input and the bottom is the synthesized result.

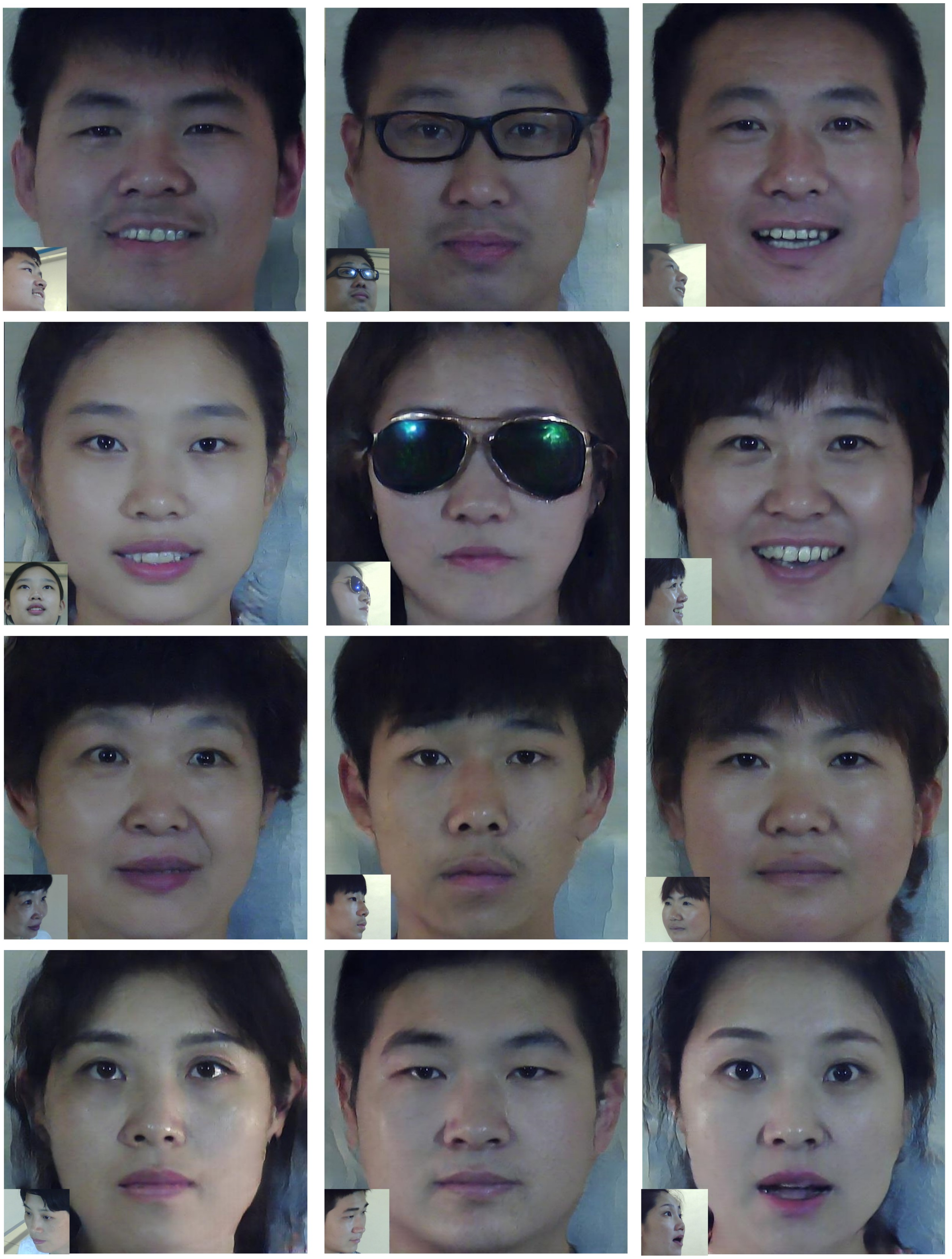

Figure 6. The 512×512 frontalization results of our method under extreme poses on M^{2}FPA. For each subject, the bottom left corner is the input image.