Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis
Supplementary Material

Rui Huang[1,2*] Shu Zhang[1,2,3*] Tianyu Li[1,2] Ran He[1,2,3]
[1]National Laboratory of Pattern Recognition, CASIA
[2]Center for Research on Intelligent Perception and Computing, CASIA
[3]University of Chinese Academy of Sciences, Beijing, China
huangrui@cmu.edu, tianyu.lizard@gmail.com, {shu.zhang, rhe}@nlpr.ia.ac.cn

1. Detailed Network Architecture

The detailed structures of the global pathway $G_{\theta^g_E}$ and $G_{\theta^g_D}$ are provided in Table 1 and Table 2. Each convolution layer of $G_{\theta^g_E}$ is followed by one residual block [2]. Particularly, the layer $conv4$ is followed by four blocks. The output of the layer $fc2$ (v_{id}) is obtained by selecting the maximum element from the two split halves of $fc1$.

The Decoder of the global pathway $G_{\theta^g_D}$ contains two parts. The first part is a simple deconvolution stack for up-sampling the concatenation of the feature vector v_{id} and the random noise vector z. The second part is the main deconvolution stack for reconstruction. Each layer takes the output of its previous layer as the regular input, which is omitted in the table for readability. Any extra inputs are specified in the $Input$ column. Particularly, the layers $feat8$ and $deconv0$ have their complete inputs specified. Those extra inputs instantiate the skipping layers and the bridge between the two pathways. The fused feature tensor from the local pathway is denoted as $local$ in Table 2. Tensor $local$ is the fusion of the outputs of four $G_{\theta^l_D}$s' layer $conv4$ (of Table 3). To mix the information of the various inputs, all extra inputs pass through one or two residual blocks before being concatenated for deconvolution. The profile image I^P is resized to the corresponding resolution and provides a shortcut access to the original texture for $G_{\theta^g_D}$.

Table 3 shows the structures of the local pathway $G_{\theta^l_E}$ and $G_{\theta^l_D}$. The local pathway contains three down-sampling and up-sampling processes respectively. The w and h denote the width and the height of the cropped patch. For the patches of the two eyes, we set w and h as 40; for the patch of the nose, we set w as 40 and h as 32; for the patch of the mouth, we set w and h as 48 and 32 respectively.

We use rectified linear units (ReLU) [4] as the non-linearity activation and adopt batch normalization [3] except for the last layer. In $G_{\theta^g_E}$ and $G_{\theta^l_E}$, the leaky ReLU is adopted.

Discussion: Our model is simple while achieving better performance in terms of the photorealism of synthesized images. Yim *et al.* [5] and Zhu *et al.* [7] use locally connected convolutional layers for feature extraction and fully connected layer for synthesis. We use weight-sharing con-

Table 1. Structure of the Encoder of the global pathway $G_{\theta^g_E}$

Layer	Filter Size	Output Size
conv0	$7 \times 7/1$	$128 \times 128 \times 64$
conv1	$5 \times 5/2$	$64 \times 64 \times 64$
conv2	$3 \times 3/2$	$32 \times 32 \times 128$
conv3	$3 \times 3/2$	$16 \times 16 \times 256$
conv4	$3 \times 3/2$	$8 \times 8 \times 512$
fc1	-	512
fc2	-	256

Table 2. Structure of the Decoder of the global pathway $G_{\theta^g_D}$. The $conv$s in $Input$ column refer to those in Table 1.

Layer	Input	Filter Size	Output Size
feat8	fc2, z	-	$8 \times 8 \times 64$
feat32	-	$3 \times 3/4$	$32 \times 32 \times 32$
feat64	-	$3 \times 3/2$	$64 \times 64 \times 16$
feat32	-	$3 \times 3/2$	$128 \times 128 \times 8$
deconv0	feat8, conv4	$3 \times 3/2$	$16 \times 16 \times 512$
deconv1	conv3	$3 \times 3/2$	$32 \times 32 \times 256$
deconv2	feat32, conv2, I^P	$3 \times 3/2$	$64 \times 64 \times 128$
deconv3	feat64, conv1, I^P	$3 \times 3/2$	$128 \times 128 \times 64$
conv5	feat128, conv0, $local$, I^P	$5 \times 5/1$	$128 \times 128 \times 64$
conv6	-	$3 \times 3/1$	$128 \times 128 \times 32$
conv7	-	$3 \times 3/1$	$128 \times 128 \times 3$

Table 3. Structure of the local pathway $G_{\theta^l_E}$ & $G_{\theta^l_D}$. The $conv$s in $Input$ column refer to those in the same table.

Layer	Input	Filter Size	Output Size
conv0	-	$3 \times 3/1$	$w \times h \times 64$
conv1	-	$3 \times 3/2$	$w/2 \times h/2 \times 128$
conv2	-	$3 \times 3/2$	$w/4 \times h/4 \times 256$
conv3	-	$3 \times 3/2$	$w/8 \times h/8 \times 512$
deconv0	conv3	$3 \times 3/2$	$w/4 \times h/4 \times 256$
deconv1	conv2	$3 \times 3/2$	$w/2 \times h/2 \times 128$
deconv2	conv1	$3 \times 3/2$	$w \times h \times 64$
conv4	conv0	$3 \times 3/1$	$w \times h \times 64$
conv5	-	$3 \times 3/1$	$w \times h \times 3$

*These two authors contributed equally

Figure 1. Our synthesized images present moderately better exposure in some cases. Each tuple consists of three images, with the input I^P on the left, the synthesized in the middle, the ground truth frontal face I^{gt} on the right. Each I^P and its corresponding I^{gt} are taken under a flash light from the same direction.

Figure 2. Synthesis results under various illuminations. The first row is the synthesized image, the second row is the input. Please to refer to the supplementary material for more results.

volution in most cases. Our model reduces parameter numbers to a large extent and avoids expensive computation for generating every pixel during synthesis. Yim *et al*. [5] and Amir *et al*. [1] add a second reconstruction branch or a refinement network. Our early supervised decoder achieves end-to-end generation of high-resolution image.

2. Additional Synthesis Results

Additional synthesized images I^{pred} are shown in Fig. 1 and Fig. 2. Under extreme illumination condition, the exposure of I^{pred} is consistent with or moderately better than that of its input I^P or its ground truth frontal face I^{gt}. Fig. 2 demonstrates TP-GAN's robustness to illumination changes. Despite extreme illumination variations, the skin tone, global structure and local details are consistent across illuminations. Our method can automatically adjust I^P's exposure and white balance.

Additionally, we use a state-of-the-art face alignment method [6] to provide four landmarks for TP-GAN under extreme poses. The result is only slightly worse than that reported in Table 2 of the paper. Specifically, we achieve Rank-1 recognition rates of 87.63($\pm$60°), 76.69($\pm$75°), 62.43($\pm$90°).

3. Activation Maps Visualization

In this part, we visualize the intermediate feature maps to gain some insights into the processing mechanism of the two-pathway network. Fig. 3 illustrates the fusion of global and local information before the final output. C_g contains the up-sampled outputs of the global pathway and C_l refers to the features maps fused from the four local pathways. Their information is concatenated and further integrated by the following convolutional layers.

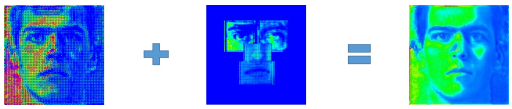

Figure 3. Synthesis process illustrated from the perspective of activation maps. The up-sampled feature map C_g is combined with the local pathway feature map C_l to produce feature maps with detailed texture.

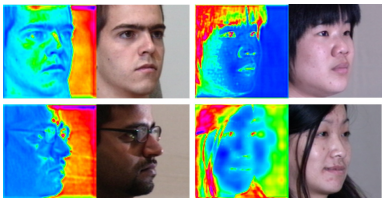

Figure 4. Automatic detection of certain semantic regions. Some skip layers' activation maps are sensitive to certain semantic regions. One for detecting non-face region is shown on the left, another for detecting hair region is shown on the right. Note the delicate and complex region boundaries around the eyeglasses and the fringe.

We also discovered that TP-GAN can automatically detect certain semantic regions. Fig. 4 shows that certain skip layers have high activation for regions such as non-face region and hair region. The detection is learned by the network without supervision. Intuitively, dividing the input image into different semantic regions simplifies the following composition or synthesis of the frontal face.

References

[1] A. Ghodrati, X. Jia, M. Pedersoli, and T. Tuytelaars. Towards automatic image editing: Learning to see another you. In *BMVC*, 2016. 2

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 1

[3] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015. 1

[4] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010. 1

[5] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015. 1, 2

[6] H. Zhang, Q. Li, and Z. Sun. Combining data-driven and model-driven methods for robust facial landmark detection. *arXiv:1611.10152*, 2016. 2

[7] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013. 1