Towards Large-Pose Face Frontalization in the Wild Supplementary Materials

Xi Yin[†]Xiang Yu[‡], Kihyuk Sohn[‡], Xiaoming Liu[†] and Manmohan Chandraker[§‡]
[†]Michigan State University
[§]University of California, San Diego
[‡] NEC Laboratories America
{yinxi1,liuxm}@cse.msu.edu, {xiangyu,ksohn,manu}@nec-labs.com

1. Network Structures

Figure 1 shows the network structure of FF-GAN, composed of the 3DMM reconstruction module R, the generator G, the discriminator D and the recognition engine C.

The 3DMM module R takes the input image $\mathbf{x}$ and generates the 3DMM coefficients $\mathbf{p}$ including weak perspective matrix $\mathbf{m} \in \mathbb{R}^{8\times1}$, shape coefficients $\alpha_{id} \in \mathbb{R}^{199\times1}$, expression coefficients $\alpha_{exp} \in \mathbb{R}^{29\times1}$, and texture coefficients $\alpha_{tex} \in \mathbb{R}^{40\times1}$. We use the provided coefficients in [3] as our ground truth for training. Originally, 3DMM consists of 199 bases for texture model. Only the first 40 bases are used in [3]. We use the CASIA-Net [2] structure, where we separate texture coefficients from shape-related coefficients in the later layers, which empirically demonstrates better performance in our experiments.

The generator G takes the image $\mathbf{x}$ and the estimated 3DMM coefficients $\mathbf{p}$ as the inputs to generate a frontal-view face $\mathbf{x}^f$. The 3DMM coefficients provide a frontal low frequency basis and the detailed appearance is expected to be recovered from the raw pose-variant input image. Clearly, these two inputs are not in the same domain. We apply three fully convolutional layers to up-sample $\mathbf{p}$ and one convolutional layer to down-sample $\mathbf{x}$ to the same size of $50\times50\times64$. The outputs are concatenated to an encoder-decoder structured network which includes two skip connections that are used to provide high frequency information to the decoding process. The feature after encoding is of dimension $512 \times 12 \times 12$ which maintains the spatial information to recover the input image.

The discriminator D aims to distinguish between the generated image $\mathbf{x}^f$ and the real frontal-view face $\mathbf{x}^g$. This is a relatively easy task, so we use a shallow network with five convolutional layers and one linear layer, which outputs a 2D vector with each dimension indicating the probability of the input belonging to the generated image or the real image. In each iteration during training, D is updated with two batches of samples from $\mathbf{x}^f$ and $\mathbf{x}^g$, respectively.

The recognition engine C also adopts a CASIA-Net structure. Instead of using the max pooling layer as CASIA-Net, we choose volumetric max pooling, which applies pooling not only in the spatial dimensions but also across the feature channels. We find this to be helpful for face recognition. C is pre-trained with CASIA-Webface dataset and fixed in the first two stages of the training process. Later, we update C using the original input image $\mathbf{x}$. Note that $\mathbf{x}^f$ are the input to fool C during the training of G and gradients flow through C to update the generator G.

2. Training Details

2.1. Further Implementation Details

For in-the-wild experiments, we train our model using 300W-LP, which is generated using the face-profiling algorithm of [3]. We prepare the training image pairs by setting one pose-variant face image (15°-90°) as the input and the frontal-view face image of the same subject (0°-15°) as the target. We use Adam solver for optimization with a batch size 128. The weight decay is set to $2e{-}4$ and momentum is set to 0.9. The initial learning rate is set to $2e{-}4$. We reduce the learning rate by a factor of 10 for every 20 epochs. As indicated in the main submission, there are three stages for training. For the first stage, λ_{rec} and λ_{id} are tuned to be 0 and 0.01, since in the early stage, the network is learning to rotate from a pose-variant face to a frontal face and the reconstruction loss or identity loss may prevent such a process. After 20 epochs, the second stage is to finetune the frontalization framework. We change λ_{rec} and λ_{id} to 1 while tuning λ_{tv} to 0.5 and λ_{sym} to 0.8. Later, when the four modules are well pre-trained, we relax the update of module C. The learning rate is $1e{-}6$ for jointly finetuning the overall framework.

For controlled experiments on Multi-PIE, we finetune from the models trained on 300W-LP. We mix the dataset of 300W-LP with Multi-PIE where 300W-LP is only used to update module R. The weights for each loss are set to 1. Since we already have a good starting point, we do not need to adjust the weights dynamically on Multi-PIE. The

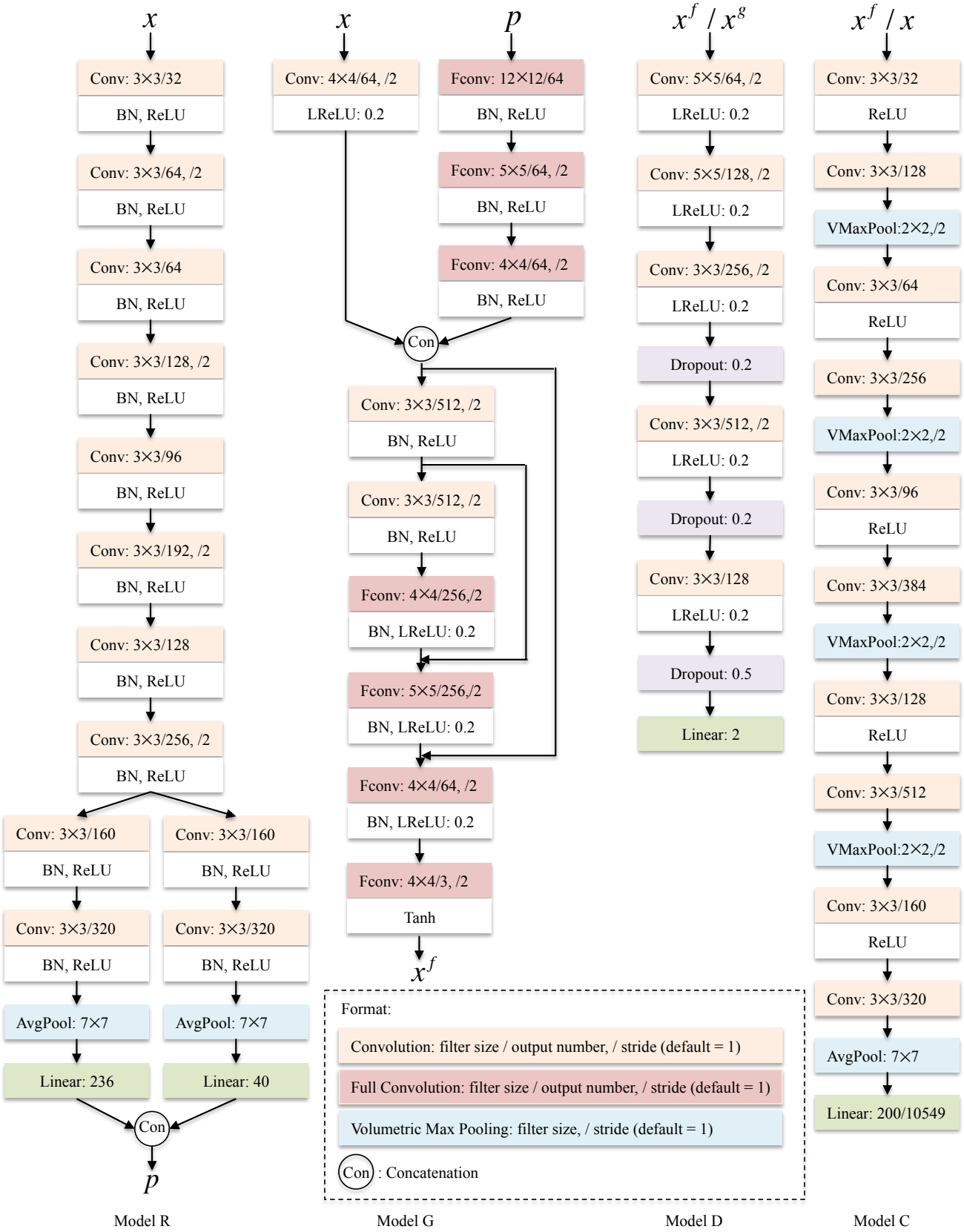

Figure 1: Network structure of FF-GAN.

Figure 2: Intermediate face frontalization results showing the three stages of the training on Multi-PIE. The beginning to epoch 5 shows the first stage, which is to rotate the face from non-frontal to frontal. The second stage lasts from epoch 5 to epoch 10, which aims to capture more fine features of the faces and collect identity information. The last stage starts from epoch 10, during which the identity information is fully recovered.

initial learning rate is set to $1e{-}4$ for the first two stages when model C is fixed and reduced to $5e{-}5$ when model C is relaxed. The first two stages need approximately 10 epochs for finetuning. The other hyper-parameters are the same as the experiments on 300W-LP.

2.2. Training Process

Figure 2 shows the training process of our face frontalization framework on Multi-PIE. We verify that during the early stages, the major task is to rotate the non-frontal face into a near-frontal pose, as shown in Figure 2, epoch 5. We illustrate training on Multi-PIE as an example here, while noting that training of 300W-LP exhibits similar trends.

After 5 epochs, our models can generate frontal-view faces, though still with artifacts or blurry effects. As the training progresses, the task becomes preserving local high frequency appearance details and identity information. From the visual results varying over the course of epoch 5 to epoch 17, we observe that local features are more and more finely captured. For example, the eyes and eye corners are generated to increasingly sharp levels of detail. At the last stage, the discriminator D usually achieves equal error for both real and generated samples, which indicates that D and G reach a balance where G can generate frontal-view faces that are realistic enough to fool D.

3. Face Frontalization Results

In this section, we will illustrate further face frontalization results on Multi-PIE, LFW, AFLW, and IJB-A datasets.

Figure 3 shows the face frontalization results of eight subjects in the test set of Multi-PIE. The proposed FF-GAN generates realistic frontal faces that are similar to the ground truth (odd rows are the input, where the frontal ground truth is the image in the middle column of odd rows) across all different poses. Furthermore, the gender, race and attributes like eyeglasses are well-preserved. It is clear that the larger the pose angle is, the more difficult it is for the generated output to preserve identity. Surprisingly, for large poses (up to $90°$), FF-GAN can still preserve the identity to a large extent. To the best of our knowledge, this is the first work to show face frontalization results for faces beyond $60°$.

Figure 4 shows the face frontalization results on LFW. Compared to previous works [1] and [4], the proposed FF-GAN can generate more realistic frontal faces in various poses and expressions. The facial detail filling technique proposed in [4] relies on a symmetry assumption and may lead to inferior results (2nd row, 6th column). In contrast, we introduce a symmetry loss in the training process that generalizes to test images without the need for post-processing to impose symmetry as a hard constraint.

Figure 3: Visual results on Multi-PIE. Each example shows 13 pose-variant inputs (Odd) and the generated frontal outputs (Even). We clearly observe that the outputs consistently recover similar frontal faces across all the pose intervals.

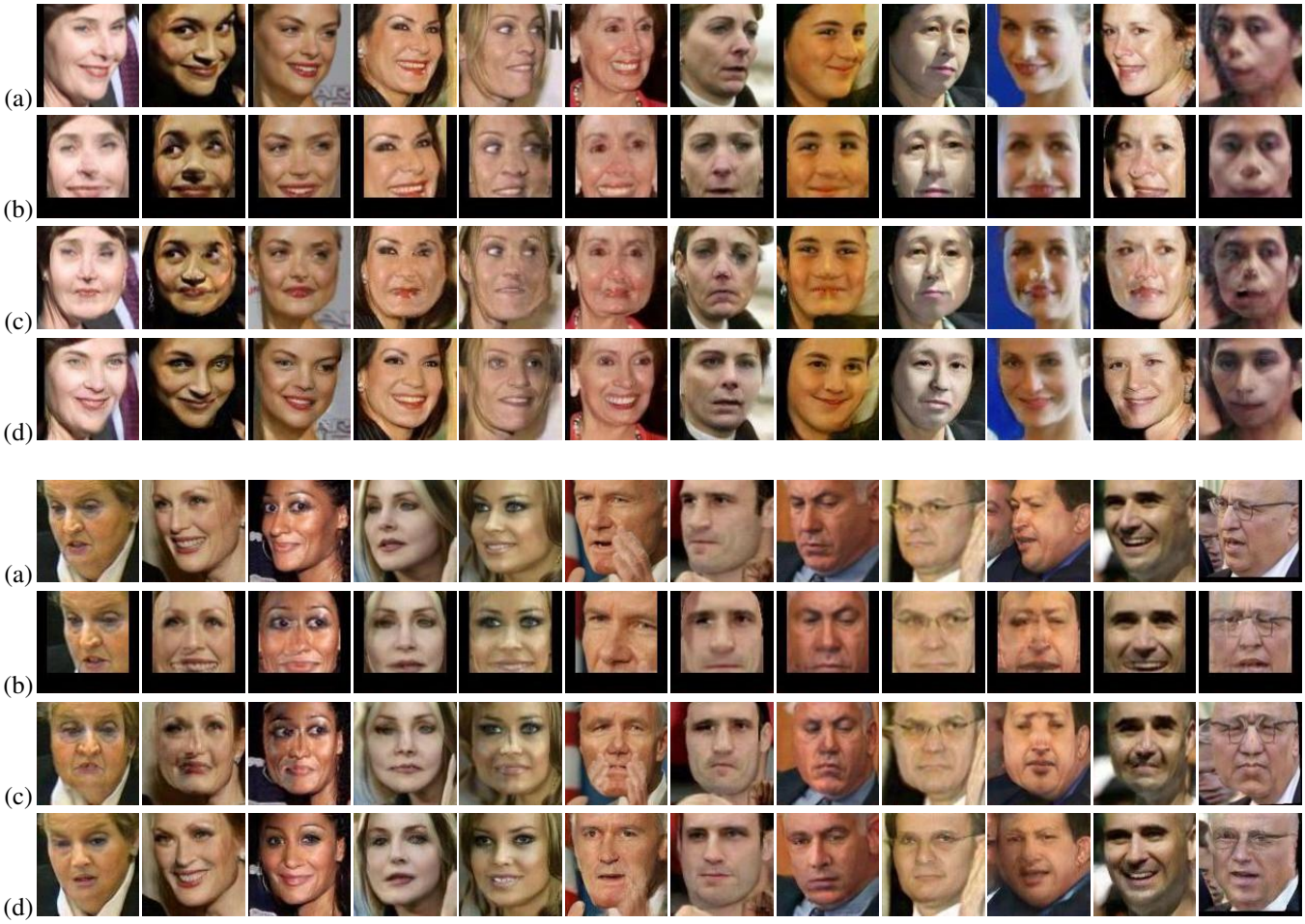

Figure 4: Face frontalization visual comparisons on LFW. (a) Input; (b) from the method of LFW-3D [1] ; (c) from the method of HPEN [4]; (d) FF-GAN (ours). We observe that our method achieves frontalizations that are much more realistic than prior works, by both recovering fine details and preserving identity.

Figure 5 shows the face frontalization results on AFLW, which encompasses more pose variations than LFW. For better visualization, we separate the faces into three different groups with small, medium and large pose variations, which are defined based on the visibility of the two eyes (both visible for small pose, one eye half-occluded for medium pose and one eye fully-occluded for large pose). FF-GAN works extremely well for the face images with small pose, in rows (a) and (b). For face images with medium or large poses in rows (c) and (d), respectively, FF-GAN still generates plausible results without many artifacts. We note that even for nearly profile views in row (d), high-frequency details of facial features are recovered well, the frontalized face is symmetric and identity is preserved quite well. Row (e) shows results for input images under various lighting or expressions. Again FF-GAN works well under these variations.

Figure 6 shows the face frontalization results on IJB-A, which consists of large-pose and low-quality face images. The input images are of medium to large pose and under a large variation of race, age, expression, and lighting conditions. However, FF-GAN can still generate realistic and identity-preserved frontal faces.

References

[1] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2014. 3, 5

[2] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint:1411.7923*, 2014. 1

[3] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 1

[4] 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. 3, 5

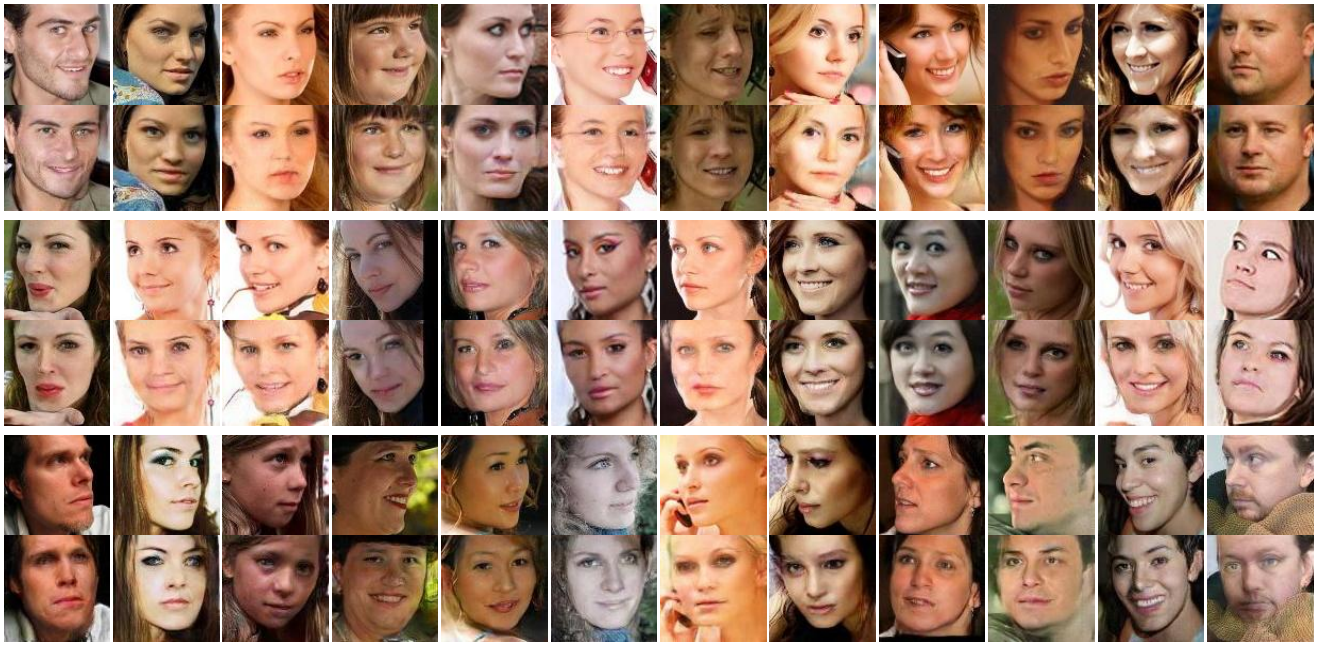

Figure 5: Face frontalization results on AFLW. Odd rows are all profile-view inputs and even rows are the frontalized results.

Figure 6: Face frontalization results on IJB-A. Odd rows are all profile-view inputs and even rows are the frontalized results.