GDFace: Gated Deformation for Multi-view Face Image Synthesis

Xuemiao Xu,[1,2,3] Keke Li,[1] Cheng Xu,[1] Shengfeng He[1]*

[1]School of Computer Science and Engineering, South China University of Technology, China
[2]State Key Laboratory of Subtropical Building Science
[3]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information

Abstract

Photorealistic multi-view face synthesis from a single image is an important but challenging problem. Existing methods mainly learn a texture mapping model from the source face to the target face. However, they fail to consider the internal deformation caused by the change of poses, leading to the unsatisfactory synthesized results for large pose variations. In this paper, we propose a Gated Deformable Face Synthesis Network to model the deformation of faces that aids the synthesis of the target face image. Specifically, we propose a dual network that consists of two modules. The first module estimates the deformation of two views in the form of convolution offsets according to the input and target poses. The second one, on the other hand, leverages the predicted deformation offsets to create the target face image. In this way, pose changes are explicitly modeled in the face generator to cope with geometric transformation, by adaptively focusing on pertinent regions of the source image. To compensate offset estimation errors, we introduce a soft-gating mechanism that enables adaptive fusion between deformable features and primitive features. Extensive experimental results on five widely-used benchmarks show that our approach performs favorably against the state-of-the-arts on multi-view face synthesis, especially for large pose changes.

Introduction

Synthesizing face images with different views has many practical applications, *e.g.*, surveillance, virtual reality, and image editing/enhancement. It is ill-posed to synthesize a face with a different view while keeping its identity well-preserved. Conventional methods resolve this problem by fitting a 3D face model from a single-view 2D face image (Hassner et al. 2015; Zhu et al. 2015). Although accurate face structure can be captured, these methods cannot "generate" occluded face regions, and therefore can only rotate the face in a small range. Deep networks, especially GAN (Goodfellow et al. 2014), show great performance in creating multi-view faces (Tran, Yin, and Liu 2017; Huang et al. 2017; Hu et al. 2018; Tian et al. 2018; Yin et al. 2017). These methods mostly adopt an encoder-decoder structure to learn the texture mapping from the input image to the target image, usually conditioned by the pose information (*e.g.*, a pose vector or two face landmark masks).

Figure 1: Existing methods fail to handle large pose variations. We tailor a dual network to cope with face deformation and synthesis simultaneously, leading to photorealistic face synthesis.

Notwithstanding the demonstrated success, large pose variation is still the main barrier. GAN-based methods learn a direct mapping between the source image and the target image but ignore the geometric transformations of faces across different views. The involved generator applies convolution operations over fixed geometric structures, making it difficult to model complex deformations of faces (*e.g.*, examples shown in Fig. 1). But indeed, modeling face deformation plays an important role in face synthesis. We observe that the rotation of faces from different identities share a similar geometric structure changes. This is a strong prior knowledge that can be incorporated as a guidance to model face deformation and benefits the synthesis of multi-view faces.

Based on the above observation, we aim to incorporate face deformation into multi-view face image synthesis. To this end, we propose a dual network that consists of two modules, each of which copes with deformation modeling and face synthesis, respectively. Given a source landmark and a target landmark, the proposed face deformation module learns to estimate the inherent transformation between two poses in the form of deformable convolution offset. It serves as the deformation prior for the face synthesis mod-

*Corresponding author. Email: hesfe@scut.edu.cn.

ule, which is composed by several gated deformable convolution blocks. These blocks perform deformable convolution operations on the source image, which enables free-form deformations of the sampling grid according to the face transformation. Although the learned deformation offset can capture large pose variations, it may increase the learning ambiguity. As a consequence, we introduce a soft-gating mechanism in our gated deformable convolution blocks, which allows a dynamic sampling between deformable features and primitive features. This strategy further enhances the capability of our model for modeling both large and small face variations. Extensive experiments demonstrate that the proposed model outperforms existing state-of-the-art methods in five widely-used datasets. To summarize, our contributions are threefold:

- We delve into the problem of large pose variations of face synthesis. We tailor a dual network, in which the first branch models face deformation in the form of convolution offset using two face landmarks, and the other leverages this strong prior for face synthesis.
- We propose to inject diverse and rich features representations to the network by presenting a soft-gating mechanism. It enables our model to adapt to different angles of face rotations by integrating deformable and primitive features.
- We outperform state-of-the-art methods on five widely-used benchmarks by a large margin, especially for the scenarios with large pose variations.

Related Work

Multi-view Face Synthesis. Traditional methods mostly tackle the problem of multi-view face synthesis by adopting 2D/3D local texture warping (Ferrari et al. 2016; Hassner et al. 2015; Zhu et al. 2015). For example, Hassner *et al*. (Hassner et al. 2015) employ a simple 3D approximation of faces to produce frontal-view face image. However, this kind of methods suffer from severe texture loss due to occlusion. Recently, there are many deep learning methods focusing on face frontalization and multi-view face synthesis (Huang et al. 2017; Hu et al. 2018; Zhao et al. 2018a; Tran, Yin, and Liu 2017; Yin et al. 2017; Tian et al. 2018; Zhao et al. 2018b). TP-GAN (Huang et al. 2017) proposes a two-pathway network to take both local details and global face structure into consideration. CAPG-GAN (Hu et al. 2018) employs the landmark mask to guide the rotation of faces, achieving multi-view face synthesis. CR-GAN (Tian et al. 2018) introduce a generation sideway to enhance generalization capacity of the network. Existing deep learning based methods can produce better results than traditional methods, but they suffer from large pose variations, which is mainly addressed in this paper.

Pose-invariant Face Recognition. Large pose variations lead to significant influences on face recognition performance. Ensuring the recognition accuracy under a large rotation angle of faces is a challenging problem. Part of existing methods aim to generate profiles for data augmentation or directly learn the pose-invariant features to eliminate the influences caused by poses (Zhu et al. 2014; Masi et al. 2016; Zhao et al. 2019). The others propose to synthesize a frontal-view face from a profile face so that face recognition can be performed in a constrained way (Huang et al. 2017; Hu et al. 2018; Zhao et al. 2018a; Tran, Yin, and Liu 2017). These generative methods are shown to be effective in many datasets. Our approach also belongs to this category, and face recognition accuracy is the major metric to evaluate the synthesized results.

Generative Adversarial Network. Generative Adversarial Network (GAN) (Goodfellow et al. 2014) is a powerful generative model which can generate samples similar to the specific data distribution through a min-max game between the generator and the discriminator. It has been extended to various applications. For example, Deep Convolutional GAN (DCGAN) (Radford, Metz, and Chintala 2015) first shows the huge potential of GAN in the task of image generation. Conditional GAN (cGAN) (Mirza and Osindero 2014) is proposed to incorporate condition constraints on the random noise, which guides the generative network to synthesize images of better quality. To help the GAN learn the interpretable representation, InfoGAN (Chen et al. 2016) proposes a mutual information regularizer for optimization. Pixel2Pixel (Isola et al. 2017) improves cGAN to deal with the image translation problem. To enable the usability of the unpair training data, CycleGAN (Zhu et al. 2017a) is proposed as an effective unsupervised learning model. All state-of-the-art face synthesis methods apply the discriminator to ensure the identity of the generated faces. To incorporate pose prior to guide the multi-view synthesis, our model adapt the design of the conditional generative adversarial network (Mirza and Osindero 2014).

Approach

In this section, we first present the architecture of the proposed network, then we discuss the proposed gated deformable block and soft-gating mechanism in detail.

Architecture

We aim to synthesize face images with multiple views, by learning the deformable face mapping between the source face and the target face given pose landmarks as guidance. To achieve this goal, we present a Gated Deformation Face Synthesis Network (see Fig. 2), which has a dual structure composed of a generator and two discriminators. The generator consists of a series of deformable convolution blocks, each integrates a gating mechanism. Specifically, the generator has two branches. The first branch takes the source face as input, and the second branch is fed with pose conditions (concatenated landmarks of the source face and the target face). Both the inputs of the two branches are encoded by the down-sampling convolution layers. Then the encoded features of the two branches are sent into the face synthesis module, which consists of several gated deformable convolution blocks. Each block aims to transform the pose of the input features progressively to the target pose under the guidance of the face deformation module. A soft-gating mechanism is introduced to control the weight of face deformation. Meanwhile, the entire network is trained end-to-end, and therefore both two branches can mutually update

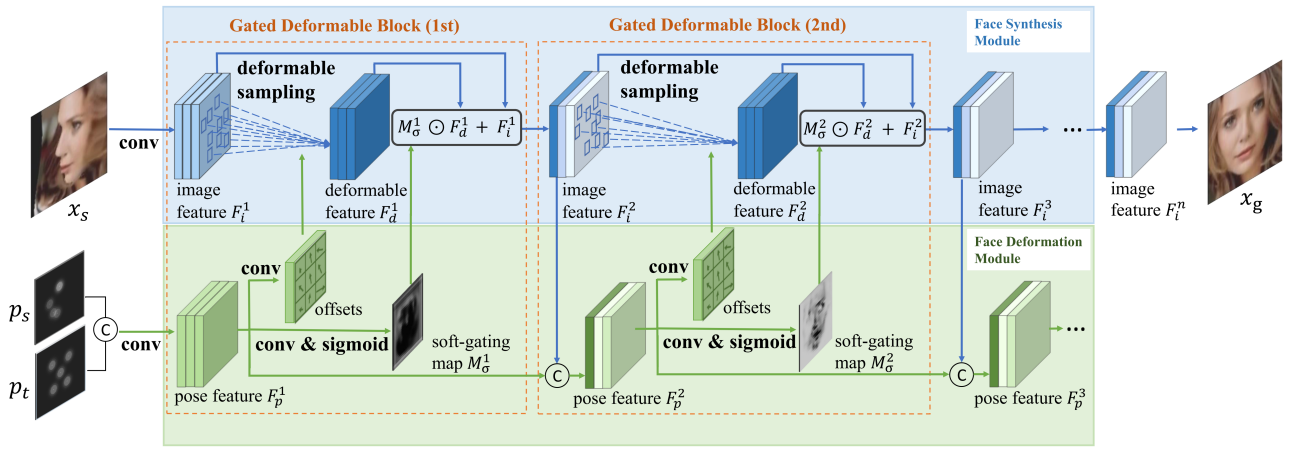

Figure 2: Overview of generator of the proposed dual network. The green region is to estimate the deformation in the form of convolution offsets from the source and target poses. The blue region takes the estimated offsets into account for deformable convolutions and generates the target face.

for producing optimized face features. The final output features in the face synthesis module are sent to a decoder for producing a face image with the target pose. Two discriminators in our model aim to provide two kinds of discriminating supervisions to ensure the consistencies of appearance (identity) and pose of the target face.

Gated Deformable Convolution

To model the face deformation caused by pose change, we introduce a Gated Deformable Convolution Block inspired by (Dai et al. 2017). As shown in Fig. 2, each block is fed with the image features F_i^t and the pose features F_p^t (t denotes the index of a block) via two branches. The image features F_i^t pass through the deformable convolution layers $Conv_d$, with the deformation offsets to get the deformable image features F_d^t. The offsets are predicted by a common convolution layer $Conv_c$ fed with the pose feature map F_p^t. For the standard 2D convolution sampling, there is a fixed grid K and corresponding weights for computing the summation of weighted sampled values. We assume $K = \{(x, y)|x, y \in \{-1, 0, 1\}\}$, which has a size of 3×3 and dilation of 1 for simplicity. We let N denote the size of K, *i.e.*, $N = |K|$. To sample a deformable feature map, we first predict the offset field $P(P \in \mathbb{R}^{H \times W \times 2N})$ by passing the pose features through $Conv_c$

$$P = Conv_c\left(F_p^t\right). \tag{1}$$

In the deformable convolution, the sampling grid K is augmented with the offsets in the grid $\{\Delta p_n|n = 1, ..., N\}$, which is reshaped from the vector $\mathbf{v}(\mathbf{v} \in \mathbb{R}^{1\times1\times2N})$ of P in the sampling grid center. For each location x in the deformable image feature map F_d^t, we could sample its value by computing

$$F_d^t(x) = \sum_{p_n \in K} w(p_n) \cdot F_i^t(x + p_n + \Delta p_n), \tag{2}$$

where w denotes the weight of convolution kernel. Since the offset Δp_n is usually fractional, we employ a bilinear interpolation kernel to compute the value of a fractional location as follows:

$$F_i^t(p) = \sum_{q \in L} H(q, p) \cdot F_i^t(q), \tag{3}$$

where p denotes a fractional location and L denotes the set of all integral neighbor locations of p. The bilinear interpolation kernel H can be formulated as

$$H(q, p) = max(0, 1 - |q_x - p_x|) \cdot max(0, 1 - |q_y - p_y|). \tag{4}$$

Once H is computed, we can get the weight of each integral neighbor location value for location p, with which we can obtain the final deformable sampling value using Eq. (3).

Our introduced gated deformable convolution block is able to conduct free-form sampling. Therefore it can not only achieve common affine transformation, but also has a strong capability to model complex face deformation caused by pose change, by sampling from the most pertinent irregular area. We give an example to demonstrate the face deformation capability of deformable convolution for multi-view face synthesis. In Fig. 3, we visualize the heat maps of offsets in each deformable convolution block. Note that, each pixel in the heat map represents the mean value of the 9 x-coordinates of the offsets in a 3×3 sampling grid since we only consider a yaw face rotation where the x-coordinates have a notable change. As shown in Fig. 3, the predicted offsets indicate the sampling direction which helps the model to "rotate" the face step by step. For the first 5 heat maps, the input profile is rotated to the frontal face, and the pixels of face are sampled towards the center of the image. For the last 6 heat maps, the face and the background are progressively sampled to the target positions. These visualization results show that the proposed gated deformable convolution can learn to model face deformation caused by large pose change. Quantitative evaluations are conducted in the experiment section.

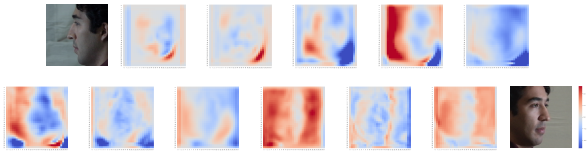

Figure 3: Deformable offsets visualization. The offset heat maps from the eleven gated deformable convolution blocks are all shown (from left to right, top to bottom). We can see that our model indeed "rotates" the source face to the target face step by step via deformable sampling.

Soft-Gating Mechanism

Although the proposed gated deformable convolution can model complex face deformation caused by pose changes, the predicted offsets inevitably introduce some estimation errors into the deformable features. In case of small pose changes, features that produced by a fixed grid structure are easy to learn, and thus may contain optimum representations. To adaptively control the weight of face deformation according to different pose changes and compensate the estimation errors introduced by offset prediction, we employ a soft-gating mechanism in the fusion of the deformable face features F_d^t of block t and the primitive face features F_i^t to get updated image features F_i^{t+1}, which will be fed into the $(t+1)$-th block. In particular, we pass the pose feature F_p^t in the pose feature branch through a convolution layer denoted by $Conv_m$, followed by a sigmoid operation over each location on the feature map. The whole process is formulated as follows:

$$M_\sigma^t = Sigmoid(Conv_m(F_p^t)), \tag{5}$$

where M_σ^t is a soft-gating map with each element ranges in [0, 1]. Then we can get weighted-fusion features F_i^{t+1} of the primitive features and the deformable features as follows:

$$F_i^{t+1} = F_i^t + M_\sigma^t \odot F_d^t. \tag{6}$$

By adjusting the values of M_σ^t, we could dynamically balance the effects of deformable features and primitive features according to their importance. In other words, the soft-gating mechanism enhances the robustness of our model to different pose variations and estimation errors, which further improves the quality of the synthesized target face.

Training Objective

To synthesize photorealistic and identity-preserved multi-view faces, we apply five different losses, including adversarial loss, pixel-wise loss, identity preserving loss, 3D shape loss , and total variation loss, to govern the training in an end-to-end manner.

Adversarial Loss. We employ two discriminators for two different objectives. Both the discriminators have two down-sampling convolution layers and several residual blocks. Discriminator D_I is responsible for keeping the identity of the generated face x_g consistent with the source face x_s. D_I takes the source face x_s and the generated face x_g as a fake pair, while the source face and the ground truth target face x_t as a real pair. Discriminator D_P has a similar structure with D_I, which aims to measure the pose consistency of x_g and the target pose p_t. D_P takes the target pose landmark p_t and the generated face x_g as a fake pair, while p_t and the target ground truth image x_t as a real pair. Under the adversarial supervision of the coupled discriminators, the generator can produce photorealistic and identity-preserved faces with specific poses. The adversarial loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{x_g\sim P_z, x_s, x_t\sim P_{data}}\left[\log D_I\left(x_t, x_s\right)\right.\\ &\left.+\log\left[1-D_I\left(x_g, x_s\right)\right]\right]\\ &+\mathbb{E}_{x_g\sim P_z, x_s\sim P_{data}, p_t\sim P_{p_t}}\left[\log D_P\left(p_t, x_t\right)\right.\\ &\left.+\log\left[1-D_P\left(p_t, x_g\right)\right]\right],\end{aligned} \tag{7}$$

where P_z, P_{data}, and P_{p_t} denote the distribution of generated faces, ground truth target faces and ground truth target landmark pose, respectively.

Pixel-wise Loss. To maintain the content consistency, we employ a pixel-wise loss:

$$\mathcal{L}_{pixel} = \sum_{i=1}^{W}\sum_{j=1}^{H}\left|x_g^{i,j} - x_t^{i,j}\right|, \tag{8}$$

where W and H denote the width and height of the image.

Identity-preserving Loss. We define the identity preserving loss as follows:

$$\mathcal{L}_{idt} = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left|F\left(x_g\right)_{i,j} - F\left(x_t\right)_{i,j}\right|, \tag{9}$$

where F denotes the feature map of the last pooling layer in a pretrained Light-CNN. The identity preserving loss forces the synthesized face to share a small distance with the ground truth image in the feature space, which enables the identity preservation when synthesizing a target-pose face.

3D Shape Loss. To further improve the quality of the synthesized face, we add a constraint that we should be able to fit a 3D face from the synthesized face which is close to the ground truth 3D face. To this end, we introduce the UV position map (Feng et al. 2018), which is a 2D image that records the 3D positions of all points in UV space to provide $\mathcal{L}_1$-like supervision $\mathcal{L}_{3D}$ on the shape and pose of face. By adding a UV position map prediction task on top of the decoder in the generator, this 3D shape guidance can improve the shape and pose consistency between the ground truth and the synthesized face. All the ground truth UV position maps are obtained by an off-the-shelf 3D face reconstruction model. We compute the 3D shape loss as follows:

$$\mathcal{L}_{3D} = \sum_{i=1}^{W}\sum_{j=1}^{H}\left|UV_g^{i,j} - UV_t^{i,j}\right|, \tag{10}$$

where W and H denote the width and height of the UV position maps. UV_g and UV_t denote the predicted and the ground truth UV position map, respectively.

Total Variation Loss. To alleviate the artifacts and obtain a smooth face image, we employ the total variation loss $\mathcal{L}_{tv}$ as used in (Huang et al. 2017).

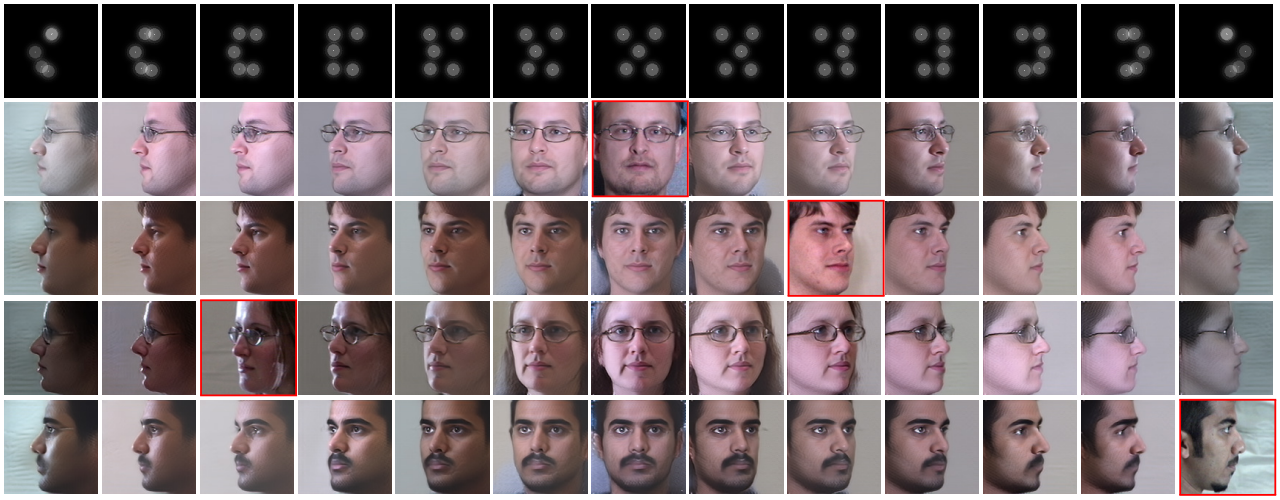

Figure 4: Multi-view face synthesis results on Multi-PIE. The first row are 13 target landmarks from $90°$ to $-90°$ with $15°$ intervals. The others are the corresponding synthesized results (inputs are marked in red).

Final Objective. The final loss of the proposed model is a weighted sum of the above losses, the generator G and two discriminators D_I and D_P are trained to alternatively optimize the min-max problem as follows:

$$\min_{G} \max_{D_I, D_P} \mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{pixel} + \lambda_3 \mathcal{L}_{idt} + \lambda_4 \mathcal{L}_{3D} + \lambda_5 \mathcal{L}_{tv}, \quad (11)$$

where λ_1, λ_2, λ_3, λ_4, λ_5 are weighted parameters to balance the above five loss items.

Experiment

In this section, we conduct extensive experiments on five widely-used datasets to demonstrate the effectiveness of our proposed model.

Datasets and Settings

Multi-PIE (Gross et al. 2010) is the largest multi-view face recognition benchmark in the constrained setting. It contains 754,204 images of 337 identities in 15 poses and 20 illumination conditions. We follow (Hu et al. 2018; Tran, Yin, and Liu 2017) to use the images from 13 poses between $-90°$ and $90°$ and 20 illuminations with the neutral expression for experiments on two different settings. The first setting only uses faces of the first 150 subjects for training and the rest 100 subjects for testing. For testing, the face with neutral expression and illumination of each subject is selected to be the gallery and all the rest faces are probes. In the second setting, all the images from 337 subjects are included. The first 200 subjects are used for training and the rest 137 subjects for testing. Each subject for testing has one gallery image with the neutral expression and illumination from its first appearance.

CelebA (Liu et al. 2015) is a large-scale face dataset in the wild, including 202,599 face images of 10,177 identities with various poses, expressions, and occlusions. We use it to demonstrate the face synthesis capability of our model on unconstrained faces.

IJB-A (Klare et al. 2015) is a challenging face dataset for face detection and recognition in the wild, it contains 5712 images and 2085 videos from 500 subjects with large variations in expressions, poses and image quality. We leverage IJB-A to evaluate the performance of our model on the unconstrained data.

CFP (Sengupta et al. 2016) is a widely-used dataset for large-pose face recognition, it contains 7000 images of 500 subjects, where each subject has 10 frontal and 4 profile face images with large pose variations. We use CFP for evaluating our model in large-pose face verification.

LFW (Huang et al. 2008) is the most commonly used databases for face recognition in the unconstrained environment. It contains 13223 from 5729 subjects with huge variations of expressions, poses, and occlusions, etc. We use LFW to evaluate the performance of our model trained with Multi-PIE in the unconstrained scenario. Furthermore, we also add the CelebA into our training set and employ the same testing protocols as in (Hu et al. 2018). Since CelebA has no pair data, we follow the method in (Zhu et al. 2017b) to generate profiles with different views from frontal faces. We use the PRNet (Feng et al. 2018) to generate ground truth UV position maps of images for training.

We compare to several state-of-the-art methods, including 3D-PIM (Zhao et al. 2018b), PIM (Zhao et al. 2018a), CAPG-GAN (Hu et al. 2018), CR-GAN (Tian et al. 2018), TP-GAN (Huang et al. 2017), and DR-GAN (Tran, Yin, and Liu 2017), both qualitatively and quantitatively.

Implementation Details

The training requires image pairs $\{x_s, x_t\}$ and pose pairs $\{p_s, p_t\}$. We enumerate all images in the datasets as source images x_s, and target pose is randomly chosen from 13 poses ranging from $-90°$ to $90°$. Once the target pose is determined, we pair x_s with its corresponding target image

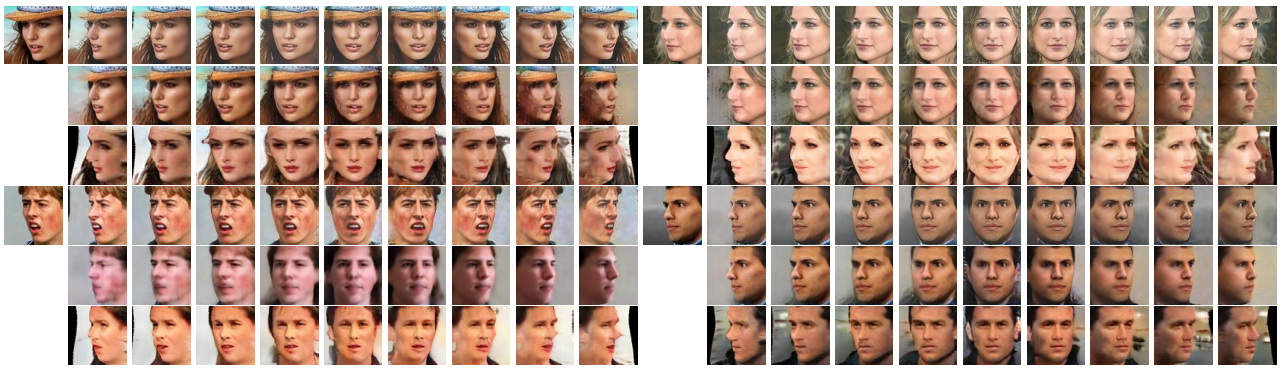

Figure 5: Multi-view synthesis results on CelebA. For each case, we show the results of ours, CAPG-GAN, and CR-GAN (top to bottom) respectively. The proposed method produces accurate and artifacts-free faces.

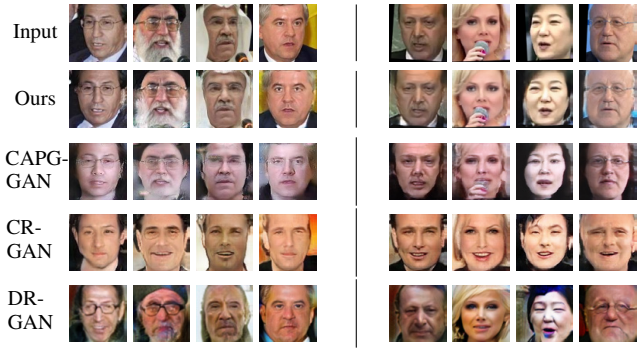

Figure 6: Visual comparison of face frontalization on the LFW (left part) and IJB-A (right part) dataset.

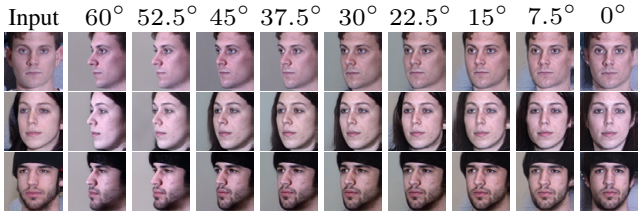

Figure 7: Synthesized faces with interpolated views on Multi-PIE.

x_t. The image size of source and target images for training is 128×128. Our network is implemented using Pytorch, the batch size is set to 16 and learning rate is 0.0001. We empirically set $\lambda_1 = 5$, $\lambda_2 = 10$, $\lambda_3 = 0.01$, $\lambda_4 = 10$, $\lambda_5 = 0.0001$. The feature extractor for identity-preserving loss is a pretrained Light CNN (Wu et al. 2018). To balance the performance and computing costs, we set the number of gated deformable convolution blocks $N_b = 11$.

We adopt 5-point facial landmarks as pose conditions, where the 5 points annotate the left eye, right eye, nose, left mouth corner, and right mouth corner, respectively. Compared with the 68-point landmark, our setting is easy to get and sufficient as a pose indicator. During training or testing,

Figure 8: Visual results of high resolution (256×256) on Multi-PIE. The first row are input faces and the second row are synthesized faces.

we first generate the landmark heatmap by plotting a Gaussian distribution centered at each keypoint on each channel, then the source and the target landmark heatmaps are sent to the model as input. During the testing phase, we use an off-the-shelf landmark detector (Zhang et al. 2016) for producing input.

Qualitative Evaluation

For face synthesis, visual effect of the synthesized image is an important metric to assess synthesizing capability of the model. We conduct experiments on Multi-PIE and CelebA to verify the superiority of our model to synthesize photo-realistic faces with multiple views. In Fig. 4, we present the multi-view synthesis results on Multi-PIE with faces from $90°$ to $-90°$ using four input faces with different poses (*i.e.*, $0°$, $-45°$, $60°$, $-90°$). It shows that our model can recover the occluded face regions even for a large pose variation and produce a photorealistic target face with a clear global structure and fine details. This attributes to the gated deformable face sampling which enables our model to learn free-from face deformation.

Meanwhile, we show the multi-view synthesis results between $60°$ and $-60°$ on CelebA in Fig. 5. The state-of-the-art methods are also presented for comparison. These synthesized faces are visual-pleasing with the identity well-preserved, which demonstrates a good generalization ability

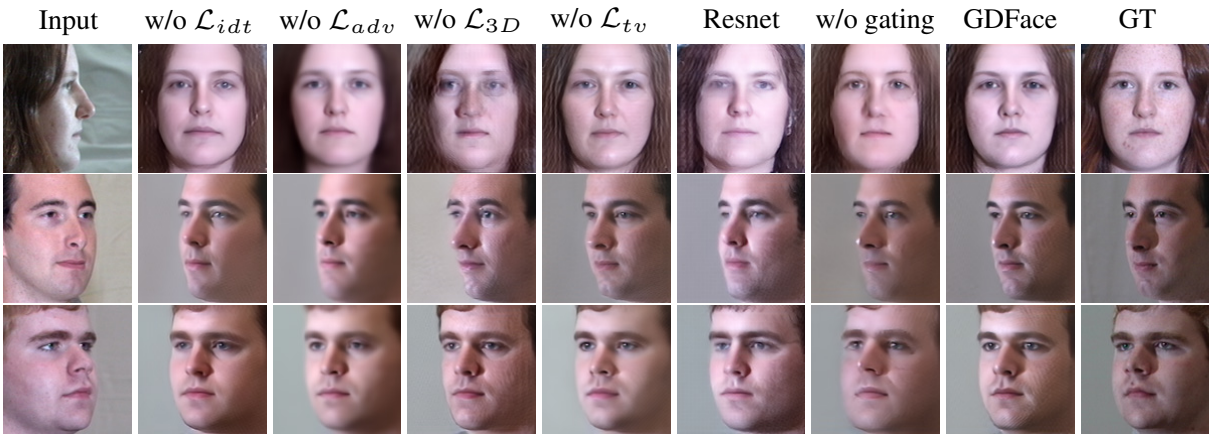

Figure 9: Visual comparison of different variants on Multi-PIE.

Methods	±90°	±75°	±60°	±45°	±30°	±15°
CPF	-	-	-	71.65	81.05	89.45
Hassner	-	-	44.81	74.68	89.59	96.78
HPN	29.82	47.57	61.24	72.77	78.26	84.23
FIP_40	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN	47.26	60.66	74.38	89.02	94.05	96.97
Light CNN	9.00	32.35	73.30	97.45	99.80	99.78
TP-GAN	64.03	84.10	92.93	98.58	99.85	99.78
PIM	75.00	91.20	97.70	98.30	99.40	99.80
3D-PIM	76.12	**94.34**	**98.84**	99.34	99.47	99.83
CAPG-GAN	77.10	87.40	93.74	98.28	99.37	99.95
Ours	**87.93**	93.74	98.28	**99.87**	**99.97**	**100**

Table 1: Rank-1 recognition rates (%) across views and illuminations under Setting 1.

Methods	±90°	±75°	±60°	±45°	±30°	±15°
FIP	-	-	45.90	64.10	80.70	90.70
CPF	-	-	61.90	79.90	88.50	95.00
DR-GAN	-	-	83.20	86.20	90.10	94.00
Light CNN	5.51	24.18	62.09	92.13	97.38	98.59
FF-GAN	61.20	77.20	85.20	89.70	92.50	94.60
TP-GAN	64.64	77.43	87.72	95.38	98.06	98.68
CAPG-GAN	66.05	83.05	90.63	97.33	99.56	99.82
PIM	86.50	95.00	98.10	98.50	99.00	99.30
3D-PIM	86.73	95.21	98.37	98.81	99.48	99.64
Ours	**90.32**	**95.28**	**98.68**	**99.68**	**99.94**	**99.97**

Table 2: Rank-1 recognition rates (%) across views and illuminations under Setting 2.

Methods	ACC(%)	AUC(%)
FF-GAN	96.42	99.45
CAPG-GAN	99.37	99.90
Ours	**99.40**	**99.90**

Table 3: Face verification accuracy (ACC) and area-under-curve (AUC) results on the LFW dataset.

of our model for the unconstrained data. On the contrary, we can see that CAPG-GAN and CR-GAN produces faces with artifacts especially on the occluded regions.

To further compare with the state-of-the-art methods, we conduct a visualization experiment on LFW and IJB-A to generate the frontal face given a profile. As shown in Fig. 6, the frontal faces generated by the competing methods fail to produce a clear global structure and recover the details which are important for identity-preserving. Besides, they contain more artifacts. By contrast, our approach can produce photorealistic face with identity well-preserved.

We also conduct experiments to demonstrate the superior synthesizing capability of our model. Fig. 7 and Fig. 8 show the synthesized faces with interpolated views and faces in high resolution, respectively. These results are visually plausible and indicate our potential for practical applications.

Quantitative Evaluation

In addition to the visual effects, we also conduct face recognition and verification experiments on four datasets to evaluate the identity-preserving capability of our model. For face recognition and verification, we leverage Light CNN (Wu et al. 2018) to extract the features of input faces and then compute the cosine similarity of the feature vectors extracted from the two faces. We evaluate the face recognition performance on Multi-PIE under two experimental settings, the numerical scores are shown in Table 1 and Table 2, respectively. For setting 1, our approach outperforms the other methods for large pose change by a large margin. For setting 2, our approach also consistently surpasses the competitors for all views. These quantitative results demonstrate that our method can preserve better identity with face rotation.

To evaluate the identity-preserving performance of our model on the unconstrained data, we further conduct face verification on LFW, CFP, and the identity similarity experiments on IJB-A following (Tian et al. 2018). In Table 3, 4, and 5, our model shows a superior performance over the other methods, which demonstrates that our model is robust to faces in the wild with various expressions, poses, and occlusions. Overall, the quantitative results prove that our model can synthesize photorealistic and identity-preserved faces in both constrained and unconstrained settings.

Methods	Similarity
DR-GAN	1.295±0.008
CR-GAN	1.217±0.010
Ours	**1.089±0.040**

Table 4: Identity similarities between real and synthesized images on the IJB-A dataset. (the lower, the better)

Method	Frontal-Profile		
	ACC	EER	AUC
Light CNN-29	92.47±1.44	8.71±1.80	**97.77±0.76**
PIM	93.10±1.01	7.69±1.29	97.65±0.62
Ours	**94.43±1.26**	**6.71±1.92**	96.59±0.91

Table 5: Face recognition performance (%) comparison on CFP. The results are averaged over 10 testing splits.

Ablation Study

Here we analyze the importance of our optimization objectives and efficacy of the main components of our model. We implement different variants of our model to analyze their performances on face recognition under setting 2 of Multi-PIE. To demonstrate the gated face deformation indeed benefits the face deformation modeling for the multi-view synthesis, we replace all the gated deformable blocks with regular residual-blocks (He et al. 2016). Meanwhile, to verify the introduced soft-gating mechanism, we remove it from our model for comparison. Table 6 reports the face recognition performance with respect to different versions of our model. The results show that each loss item we use indeed boosts the performance of our model, and our deformable convolution module significantly outperforms the regular residual-block for large pose changes. Also, with the introduced gating mechanism, our final model performs better for most angles than the other variants. This is mainly because the introduced gating mechanism injects diverse face representations to the network. The advantages brought by these two modules become larger for large pose changes, as they can control the face deformation and improve the quality of the synthesized faces. We also show the visual performance of different variants of our model in Fig. 9. When removing $\mathcal{L}_{idt}$, the synthesized faces are smooth and thus lack important discriminative details. Faces become blurry when $\mathcal{L}_{adv}$ is dropped. $\mathcal{L}_{3D}$ is important for preserving the global structure of face and the texture will be smoother when adopting $\mathcal{L}_{tv}$. Finally, gated deformable convolution enables our model to generate photorealistic faces with higher quality.

Furthermore, we conduct an experiment to explore the relationship between our gated deformable convolution and rotation degrees. We plot the mean value M_σ^{mean} of the soft-gating map M_σ when rotating faces with different poses to the frontal face to see how M_σ^{mean} changes. As shown in Fig. 10, both in the setting 1 and setting 2 on Multi-PIE, the value of M_σ^{mean} increases along with the rotation degrees. It confirms our conjecture that the network relies more on the deformable blocks for large pose variations. This also indicates the importance of the proposed deformable module,

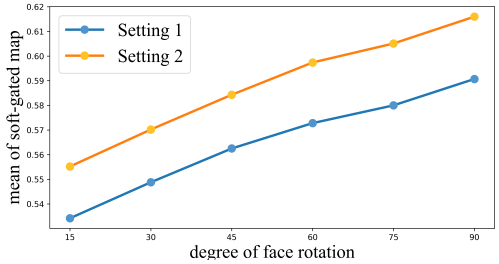

Figure 10: Tendency of the mean value of M_σ with respect to face rotation degrees. M_σ increases along with rotation degree, which demonstrates the effectiveness of our deformable module on large pose variations.

Model	±90°	±75°	±60°	±45°	±30°	±15°
w/o $\mathcal{L}_{idt}$	15.14	33.07	46.20	69.08	87.45	97.34
w/o $\mathcal{L}_{adv}$	86.76	94.10	98.30	99.80	99.94	99.98
w/o $\mathcal{L}_{3D}$	88.15	93.74	97.35	99.18	99.94	**100**
w/o $\mathcal{L}_{tv}$	89.00	94.56	98.01	99.55	99.89	99.96
Resnet	87.43	93.99	97.98	99.55	**99.97**	**99.99**
w/o gating	89.24	94.34	98.30	99.58	**99.97**	**99.99**
Ours	**90.32**	**95.28**	**98.68**	**99.68**	99.94	99.97

Table 6: Rank-1 recognition rates (%) of our model and its variants with different training objectives under Setting 2.

and the proposed gating mechanism provides a good balance between primitive features and deformable features.

Conclusion

In this paper, we propose a Gated Deformable Face Synthesis Network for multi-view face synthesis. It captures face deformation of two poses in the form of convolution offsets. This information serves as a strong prior for face synthesis via gated deformable blocks, which enables learning a complex face deformation. Furthermore, we introduce a soft-gating mechanism in each block to adaptively alleviate the estimation errors of predicted offsets and inject diversity to the feature representations. Extensive quantitative and qualitative experiments on five widely-used datasets demonstrate that the proposed method can synthesize photorealistic multi-view faces while preserving identity under both constrained and unconstrained settings, especially for large pose changes.

Acknowledgements

The work is supported by NSFC (Grant No. 61772206, U1611461, 61472145, 61702194, 61972162), Guangdong R&D Key Project of China (Grant No. 2018B010107003), Guangdong High-level Personnel of Special Support Program (Grant No. 2016TQ03X319), Guangdong Natural Science Foundation (Grant No. 2017A030311027), Guangzhou Key Project in Industrial Technology (Grant No. 201802010027, 201802010036), and the CCF-Tencent Open Research fund (CCF-Tencent RAGR20190112).

References

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2172–2180.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *ICCV*, 764–773.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 534–551.

Ferrari, C.; Lisanti, G.; Berretti, S.; and Del Bimbo, A. 2016. Effective 3d based frontalization for unconstrained face recognition. In *ICPR*, 1047–1052.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, 2672–2680.

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; and Baker, S. 2010. Multi-pie. *Image and Vision Computing* 28:807–813.

Hassner, T.; Harel, S.; Paz, E.; and Enbar, R. 2015. Effective face frontalization in unconstrained images. In *CVPR*, 4295–4304.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018. Pose-guided photorealistic face rotation. In *CVPR*, 8398–8406.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2439–2448.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Klare, B. F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; and Jain, A. K. 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 1931–1939.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Masi, I.; Rawls, S.; Medioni, G.; and Natarajan, P. 2016. Pose-aware face recognition in the wild. In *CVPR*, 4838–4846.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, 1–9.

Tian, Y.; Peng, X.; Zhao, L.; Zhang, S.; and Metaxas, D. N. 2018. Cr-gan: learning complete representations for multi-view generation. In *IJCAI*, 942–948.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 1415–1424.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security* 13:2884–2896.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2017. Towards large-pose face frontalization in the wild. In *ICCV*, 3990–3999.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23:1499–1503.

Zhao, J.; Cheng, Y.; Xu, Y.; Xiong, L.; Li, J.; Zhao, F.; Jayashree, K.; Pranata, S.; Shen, S.; Xing, J.; et al. 2018a. Towards pose invariant face recognition in the wild. In *CVPR*, 2207–2216.

Zhao, J.; Xiong, L.; Cheng, Y.; Cheng, Y.; Li, J.; Zhou, L.; Xu, Y.; Karlekar, J.; Pranata, S.; Shen, S.; et al. 2018b. 3d-aided deep pose-invariant face recognition. In *IJCAI*, 11.

Zhao, J.; Xiong, L.; Li, J.; Xing, J.; Yan, S.; and Feng, J. 2019. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE Trans. Pat. Ana. & Mach. Int.* 41:2380–2394.

Zhu, Z.; Luo, P.; Wang, X.; and Tang, X. 2014. Multi-view perceptron: a deep model for learning face identity and view representations. In *NeurIPS*, 217–225.

Zhu, X.; Lei, Z.; Yan, J.; Yi, D.; and Li, S. Z. 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 787–796.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2223–2232.

Zhu, X.; Liu, X.; Lei, Z.; and Li, S. Z. 2017b. Face alignment in full pose range: A 3d total solution. 41:78–92.