Load Balanced GANs for Multi-view Face Image Synthesis

Jie Cao[1,2,4], Yibo Hu[1,2,4], Bing Yu[5], Ran He[1,2,3,4] and Zhenan Sun[1,2,3,4]
[1]National Laboratory of Pattern Recognition, CASIA
[2]Center for Research on Intelligent Perception and Computing, CASIA
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS
[4]University of Chinese Academy of Sciences, Beijing, 100049, China
[5]Noah's Ark Lab of Huawei Technologies
{jie.cao,yibo.hu}@cripac.ia.ac.cn, yubing5@huawei.com, {rhe, znsun}@nlpr.ia.ac.cn

Abstract

Multi-view face synthesis from a single image is an ill-posed problem and often suffers from serious appearance distortion. Producing photo-realistic and identity preserving multi-view results is still a not well defined synthesis problem. This paper proposes Load Balanced Generative Adversarial Networks (LB-GAN) to precisely rotate the yaw angle of an input face image to any specified angle. LB-GAN decomposes the challenging synthesis problem into two well constrained subtasks that correspond to a face normalizer and a face editor respectively. The normalizer first frontalizes an input image, and then the editor rotates the frontalized image to a desired pose guided by a remote code. In order to generate photo-realistic local details, the normalizer and the editor are trained in a two-stage manner and regulated by a conditional self-cycle loss and an attention based L2 loss. Exhaustive experiments on controlled and uncontrolled environments demonstrate that the proposed method not only improves the visual realism of multi-view synthetic images, but also preserves identity information well.

1 Introduction

Multi-view face image synthesis has plenty of applications in various domains including pose-invariant face recognition, virtual and augmented reality, and computer graphics. Although humans can easily conceive different views of a face in mind when seeing it, making the computer have this conceive (synthesis) ability is an appealing and long-standing challenge. Traditional methods resort to 3D Morphable Model (3DMM) [Blanz and Vetter, 1999] to address this challenge. They build 3D face model as reference and then synthesize face images with new angles through model fitting. Although these 3D methods can synthesize or rotate a face image to some extent, their synthesis results are often not photo-realistic.

Recently, face synthesis models based on convolutional neural networks (CNNs) have drawn much attentions. These methods are built on black-box models and do not depend on 3D facial shape. Without explicitly modeling a face, they produce the output under the control of remote code [Ghodrati *et al.*, 2016; Yim *et al.*, 2015]. For instance, if the remote code of yaw angle is set to 30°, then the networks will automatically rotate an input image with an arbitrary pose to 30°. Recently, with the application of Generative Adversarial Networks (GAN) in multi-view face image synthesis, much progress has been made [Tran *et al.*, 2017; Zhao *et al.*, 2017]. However, when desired synthesis pose tends to be larger, it is still easy for humans to distinguish the synthesized images from the genuine ones.

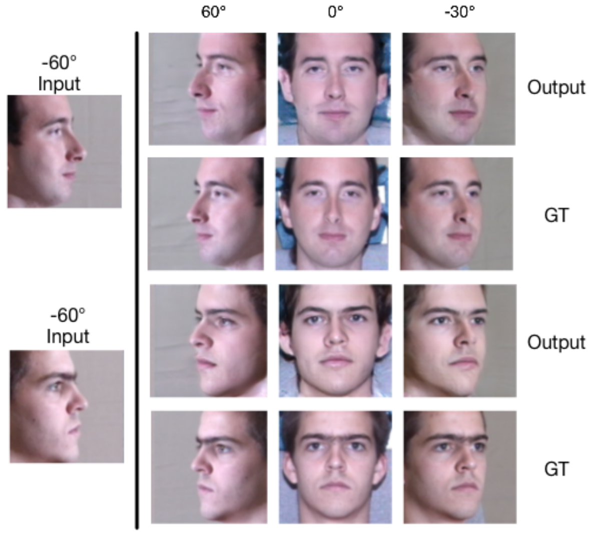

Figure 1: Face rotation results by our LB-GAN. According to the degrees on the top for each column on the right-hand, the inputs are rotated to a specified pose. GT stands for the ground truth.

This paper addresses the challenging multi-view face image synthesis problem by simplifying it into two subtasks, resulting in a new method named Load Balance Generative Adversarial Networks (LB-GAN). Concretely, we employ two pairs of GAN whose generators cooperate with each other. The first GAN consists of a generator termed as face normalizer and the corresponding discriminator. The face normalizer only focuses on frontalizing face images. The second GAN consists of another generator termed as face editor and its discriminator. The face editor takes frontal view face images as additional inputs and rotates the input image to a specified pose according to a given remote code. We combine the normalizer and the editor together to rotate the yaw angle of an input face to any specified one. Some synthesized samples by LB-GAN are shown in Fig. 1.

We employ a two-stage strategy to train LB-GAN. In the first stage, only the face normalizer and its discriminator are trained through the conventional manner for GAN [Goodfellow *et al.*, 2014]. After plausible results have been produced by the normalizer, we begin training the whole model in the second stage. Considering that noisy backgrounds of face images captured in unconstrained environments will degrade visual realism of the result severely, we propose a novel conditional self-cycle loss and an attention based $L2$ loss to tackle this problem. Experimental results on Multi-PIE and IJB-A show that our method can produce photo-realistic multi-view face images. Besides, the performance of pose-invariant face recognition is boosted through our synthetic results. In summary, the main contributions of our work are:

1) We propose LB-GAN that simplifies the ill-posed multi-view face synthesis problem into two well constrained ones.

2) Trained in a novel two-stage method, our model can preserve abundant identity information while rotating a face to arbitrary poses.

3) Profiting from the conditional self-cycle loss and attention based $L2$ loss, our model is robust to noisy environments.

4) Experimental results show that our model produces photo-realistic multi-view face images and obtains state-of-the-art cross-view face recognition performance under both controlled and uncontrolled environments.

2 Related Work

2.1 Face Frontalization

Face frontalization can be regarded as a single view face image synthesis problem, i.e., producing the frontal face images is the only consideration. To eliminate the influence of poses in face recognition or other facial analysis tasks, face frontalization has been widely studied in recent years. 3D-based models [Dovgard and Basri, 2004; Hassner, 2013; Hassner *et al.*, 2015; Zhu *et al.*, 2015; Ferrari *et al.*, 2016] are proposed for frontalization in controlled environments. Besides, deep learning models are also very competitive, e.g., CNNs [Zhu *et al.*, 2013; 2014], auto-encoders [Zhang *et al.*, 2013; Kan *et al.*, 2014] and recurrent neural networks (RNNs) [Yang *et al.*, 2015]. At present state-of-the-art face frontalization methods in controlled [Huang *et al.*, 2017] and in-the-wild [Zhao *et al.*, 2017] settings are both based on GAN.

For face frontalization and other face synthesis tasks, the capacity of identity preservation is mainly evaluated through face recognition. To this end, [Tran *et al.*, 2017] extract pose-robust identity representations from the face generator for recognition, while [Hassner *et al.*, 2015; Huang *et al.*, 2017] directly use the synthesized face images for recognition.

2.2 Generative Adversarial Networks

GAN is a novel deep framework proposed by [Goodfellow *et al.*, 2014]. GAN can be regarded as a two-player non-cooperative game model. The main components of GAN, generator and discriminator, are rivals of each other. The generator tries to map some noise distribution to the data distribution. The discriminator tries to distinguish the fake data produced by the generator from the real data. In practice, the parameters of the generator and the discriminator are trained alternately until convergence. The most significant contributions of GAN is the remarkable improvement on visual realism. Conditional GAN is proposed by [Mirza and Osindero, 2014] to send conditional information to both the generator and discriminator. To deal with unpaired data, [Zhu *et al.*, 2017] propose CycleGAN.

3 Proposed Methods

Assume there are n_{id} subjects in the training set and each face image $\mathbf{x}$ has corresponding identity label y^{id} and pose label y^p. $y^{id} \in \{1, 2, \cdots, n_{id}-1, n_{id}\}$. $y^p \in \{-90°, -75°, \cdots 0°, \cdots 90°\}$. So y^p has $n_p = 13$ discrete possible values. Remote code $\mathbf{c}$ is a n_p-dimensional one hot vector. We assign the c^*th element in $\mathbf{c}$ to 1 only if we want to change the pose of input to the c^*th type. Our goal is to train a model which takes a remote code $\mathbf{c}$ and maps the given $\mathbf{x}$ to a new face image $\hat{\mathbf{x}}$. $\hat{\mathbf{x}}$ should meet the following three requirements: (1) the visualization of $\hat{\mathbf{x}}$ is realistic, (2) the identity of $\hat{\mathbf{x}}$ remains the same as $\mathbf{x}$, (3) the pose is altered according to the specified $\mathbf{c}$.

3.1 Model Structure

As illustrated in Fig. 2, our proposed LB-GAN consists of a pair of GAN to address the multi-view face synthesis problem. The first GAN is composed of the face normalizer G_N and the corresponding discriminator D_N. Similarly, the second GAN is composed of the face editor G_E and D_E. During the test phase, G_N first takes $\mathbf{x}$ and transforms it into frontal view face image (we denote the pose label of frontal view face images as y^{p^*} below), then G_E takes $\mathbf{x}$, the output of G_N and the remote code $\mathbf{c}$ to produce the desired $\hat{\mathbf{x}}$.

In the training stage, D_N takes fake images produced by G_N or genuine images draw from datasets with pose label y^{p^*}. Similar with [Tran *et al.*, 2017], the goal of D_N is giving explicit identities of input images rather than simply judging whether they are produced by G_N. $D_N(\mathbf{x})$ is the prediction for the identity made by D_N. $D_N(\mathbf{x}) = [D_N^1(\mathbf{x}), D_N^2(\mathbf{x}), \cdots, D_N^{n_{id}}(\mathbf{x}), D_N^{y^{id^*}}(\mathbf{x})]$, where $D_N^i(\mathbf{x})$ stands for the probability that the identity label of $\mathbf{x}$ equals to i. The identity labels of produced images are all y^{id^*}. Fed by $\mathbf{x}$ with y^{id}, G_N aims to fool D_N into believing the produced image having identity label y^{id}. The objective functions of D_N and G_N can be formulated as:

$$\begin{aligned}\max_{\Theta_{D_N}} V(\Theta_{D_N}) &= \mathbb{E}_{\mathbf{x}, y^{id} \sim p_m}[\log D_N^{y^{id}}(\mathbf{x})] \\ &\quad + \mathbb{E}_{\substack{\mathbf{x}, y^p, y^{id} \\ \sim p_{data}}}[\log D_N^{y^{id^*}}(G_N(\mathbf{x}))]\end{aligned} \tag{1}$$

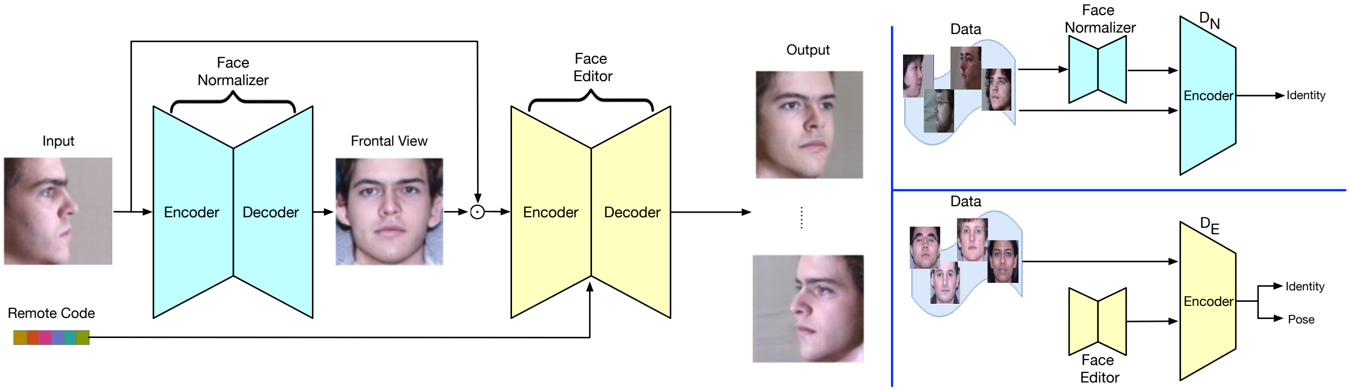

Figure 2: The framework of our LB-GAN for multi-view face image synthesis.

$$\max_{\Theta_{G_N}} V(\Theta_{G_N}) = \mathbb{E}_{\substack{\mathbf{x},y^p,y^{id}\\ \sim p_{data}}}[\log D_N^{y^{id}}(G_N(\mathbf{x}))] \tag{2}$$

where Θ_{D_N} and Θ_{G_N} denote the parameter sets of D_N and G_N respectively, $p_{data} = p_{\mathbf{x},y^p,y^{id}}(\mathbf{x}, y^p, y^{id})$, $p_m = p_{\mathbf{x},y^{id}|y^p=y^{p*}}(\mathbf{x}, y^{id})$. The first item in Eq. (1) pushes D_N to recognize the identities of the subjects in the training set, and the second one pushes D_N to find the images produced by G_N. In the meantime, to maximize Eq. (2), G_N has to keep the identity information of the input well preserved. Although D_N does not directly discriminates the pose label, G_N has to transform the pose of its input into y^{p^*} to be in accordance with genuine data. In such an adversarial training procedure, D_N will be able to distinguish the real and the fake, and G_N will be able to produce photo-realistic images.

D_E, which can be regarded as a multi-task classifier, predicts poses as well as identities. The predictions of the identity and the pose made by D_E are denoted as $D_E(\mathbf{x})$ and $D_{E_p}(\mathbf{x})$ respectively. Their definitions are similar with $D_N(\mathbf{x})$. Guided by the remote code $\mathbf{c}$, the goal of G_E is to alter the pose of input to the c^*th type without being discovered by D_E. $\mathbf{c}$ is added to G_E through the way proposed by [Salimans *et al.*, 2016]. Formally, G_E and D_E are optimized as follows:

$$\begin{aligned}\max_{\Theta_{D_E}} V(\Theta_{D_E}) = \mathbb{E}_{\substack{\mathbf{x},y^p,y^{id}\\ \sim p_{data}}}&[\log D_E^{y^{id}}(\mathbf{x}) + \log D_{E_p}^{y^p}(\mathbf{x})\\ &+ \log D_E^{y^{id^*}}(G_E(\mathbf{x}, G_N(\mathbf{x}), \mathbf{c})]\end{aligned} \tag{3}$$

$$\begin{aligned}\max_{\Theta_{G_E}} V(\Theta_{G_E}) = \mathbb{E}_{\substack{\mathbf{x},y^p,y^{id}\\ \sim p_{data}}}&[\log D_{E_p}^{c^*}(G(\mathbf{x}, G_N(\mathbf{x}), \mathbf{c}))\\ &+ \log D_E^{y^{id}}(G_E(\mathbf{x}, G_N(\mathbf{x}), \mathbf{c}))]\end{aligned} \tag{4}$$

where Θ_{D_E} and Θ_{G_E} denote the parameter sets of D_E and G_E respectively. Since D_E is trained to discriminate poses as well, G_E need to change the pose of its input but keep the visual realism and the identity information well-preserved. Note that G_E takes both $\mathbf{x}$ and the face frontalized by G_N as input. Different from previous approaches that synthesize multi-view face images by a single generator, our G_E get more information from G_N to obtain robustness of dealing with variant poses. The frontalized face will be helpful for rotating face with extreme poses, and the original input face will contribute more for identical and symmetric conditions (e.g., rotate $60°$ to $60°$ and rotate $-30°$ to $30°$).

3.2 Two-stage Training Method

The training process of our LB-GAN is two-staged. In the first stage, we only train the face normalizer and its discriminator in the alternative and adversarial manner [Goodfellow *et al.*, 2014]. We stop the process when visually appealing results have been generated by G_N. Then in the second stage, we train the whole model. We find making the parameters of G_N near-optimal first will stabilize the second training procedure and guarantee better final performance. Specifically, we optimize the parameters by Adam optimizer [Kingma and Ba, 2015] with a learning rate of 2e-4 and momentum of 0.5. The first training stage lasts for 20,000 iterations. In the second stage, the learning rate for G_N and D_N is reduced to a quarter. We train 4 iterations for optimizing G_N and G_E, and then 1 iteration for D_N and D_E. The batch size is set to 24. Note that extra regularization items, which will be discussed in section 3.3, are added in the second stage.

3.3 Regularization Items

Attention based L2 Loss. $L2$ loss is a common choice for measuring the difference of two images, and every pixel is treated equally. However, to make synthesized face images realistic and characteristic, some key facial parts should be emphasized, like eyes, mouth, nose, etc. Further, for those images captured in real life condition, the styles of clothes and hair of subjects and the background tend to change frequently. So minimizing $L2$ loss will make those regions blurry. To this end, attention based $L2$ loss denoted as $L2'$ is proposed:

$$L2'(\mathbf{x}, \hat{\mathbf{x}}) = \|(\mathbf{x} - \hat{\mathbf{x}}) \circ M\|_2 \tag{5}$$

where the operator $\circ$ denotes the Hadamard product, and M is a mask whose entries are set to 1 for the region of interest and 0 otherwise. Through adding M, our model is guided to concentrate on synthesizing convincing facial images and avoid putting too much attention on unnecessary details in the background.

Obviously, the optimal M should exactly cover the facial part of $\mathbf{x}$ and exclude the other parts. But choosing the optimal M for calculating $L2'$ loss is very expensive. Laborious manual annotation or face parsing algorithm with high accuracy is required. We sidestep this demand by loosening the restriction of M. Concretely, M only covers a few image patches that contain key parts of face. The location of the patches can be determined by the landmarks, which are also used for face image preprocessing. Since the input face image will be scaled and aligned, the proper patch sizes and locations for one image also works well for the others, we keep the the patch sizes and locations fixed for all images. The specify choice of $L2'$ will be given in the experiment section.

Conditional self-cycle loss (L_{csc}). L_{csc} is come up with such an observation: if $\mathbf{c}$ exactly matches the pose label y^p, e.g., the input $\mathbf{x}$ is a frontal view image and the remote code $\mathbf{c}$ sets the output to be frontal view as well, then $\mathbf{x}$ itself is the optimal result, so $\hat{\mathbf{x}}$ should be the same with $\mathbf{x}$ in such a case (the case is denoted as "$\mathbf{x}$ is the optimal output below"). Our L_{csc} is formulated as:

$$L_{\text{csc}} = \begin{cases} L_2'(\mathbf{x}, \hat{\mathbf{x}}), & if\ \mathbf{x} \text{ is the optimal output} \\ 0, & otherwise \end{cases} \tag{6}$$

We use $L2'$ to measure the difference between $\mathbf{x}$ and $\hat{\mathbf{x}}$. The name of L_{csc} is similar with cycle consistency loss (L_{cyc}) proposed by [Zhu *et al.*, 2017]. However, L_{cyc} is designed to enforce the networks to find the corresponding relationship when training with unpaired data and is originally proposed for domain transfer problem, such as image-to-image translation. In contrast, L_{csc} enforces the networks to avoid redundant operation on the input, i.e., the networks are encouraged to keep the input identical in some circumstance. Besides, the input and the output are in the same semantic domain in the face image synthesis problem.

4 Experiments and Analysis

4.1 Experimental Settings

Datasets. Three datasets are involved in our experiments: Multi-PIE [Gross *et al.*, 2010], IJB-A [Klare *et al.*, 2015] and CASIA-WebFace [Yi *et al.*, 2014]. Multi-PIE is established for studying on PIE (pose, illumination and expression) invariant face recognition. 20 illumination conditions, 13 poses within ±90° yaw angles and 6 expressions of 337 subjects were captured in controlled environments. IJB-A is another database with large pose variations. It has 5, 396 images and 20, 412 video frames of 500 subjects. CASIA-WebFace is a large-scale dataset containing 10,575 subjects and 494,414 images. The collection process started from the well-structured information in IMDB and then continued by using web crawler.

Table 1: Benchmark comparison of identification rate (%) across poses on Multi-PIE. Methods marked with † can only produce frontal view face images. Methods marked with ∗ are ordinary face recognition ones that are not designed for pose invariant recognition.

Method	±15°	±30°	±45°	±60°	±75°	±90°
DR-GAN	94.9	91.1	87.2	84.6	-	-
FF-GAN†	94.6	92.5	89.7	85.2	77.2	61.2
TP-GAN†	98.7	98.1	95.4	87.7	77.4	64.6
LightCNN∗	98.6	97.4	92.1	62.1	24.2	5.5
LB-GAN(Ours)	**99.1**	**98.9**	**96.7**	**91.0**	**80.3**	**65.4**

Table 2: Mean head pose estimation errors (in degree) on Multi-PIE predicted by THPE.

	±30°	±22.5°	±15°	±7.5°	0°
Genuine data	3.0	-	3.2	-	2.1
Synthesized data	4.6	5.7	4.0	5.1	2.9

For experiments on Multi-PIE, we use the first 200 subjects for training and the rest 137 for testing. Each testing identity has one gallery image from his/her first appearance. Hence, there are 72,000 and 137 images in the probe and gallery sets respectively. There are no overlap subjects between the training and testing sets. To test the performance on IJB-A, we train our model on Multi-PIE and CASIA-WebFace. We follow the testing protocol in [Tran *et al.*, 2017].

Data Preprocessing. Face images in those datasets are normalized to 96 × 96 before fed into our model. Image intensities are linearly scaled to the range of $[-1, 1]$. We use the landmarks of the centers of eyes and mouth to normalize images by the method proposed by [He *et al.*, 2017]. Note that the normalization in this step is apparently different from the function of G_N. The mask M consists of three parts: two 15 × 15 patches centered at the eyes and a 20 × 20 patch centered at the mouth center.

Implementation Details. We employ the improved version [Tran *et al.*, 2017] of CASIA-Net [Yi *et al.*, 2014] to see if our novel structure can push the limit set by them. The CASIA-Net, which can be regarded as an encoder, is able to transform an input image into an identity representation. To map the representation back to image, we build the decoder through replacing the convolution layer with transposed convolution layers. G_N and G_E have the encoder-decoder structure, D_N and D_E only have the encoder structure. Fully connected layers are added for discriminators to predict the identities and the poses. The remote code $\mathbf{c}$ is injected into the bottleneck layer of G_E.

4.2 Comparison Results

To demonstrate the effectiveness of our method, we make comparisons with several state-of-the-art ones. The performances are evaluated both qualitatively and quantitatively. Specifically, three aspects are considered: the visual quality, the performances on pose-invariant face recognition and head pose estimation.

Visual Quality. The results in Fig. 3 show how our LB-GAN rotates the face images in Multi-PIE to specific poses. The poses are not limited to the 13 discrete values because we can produce any continuous value through interpolation [Radford *et al.*, 2016]. For instance, we average the remote

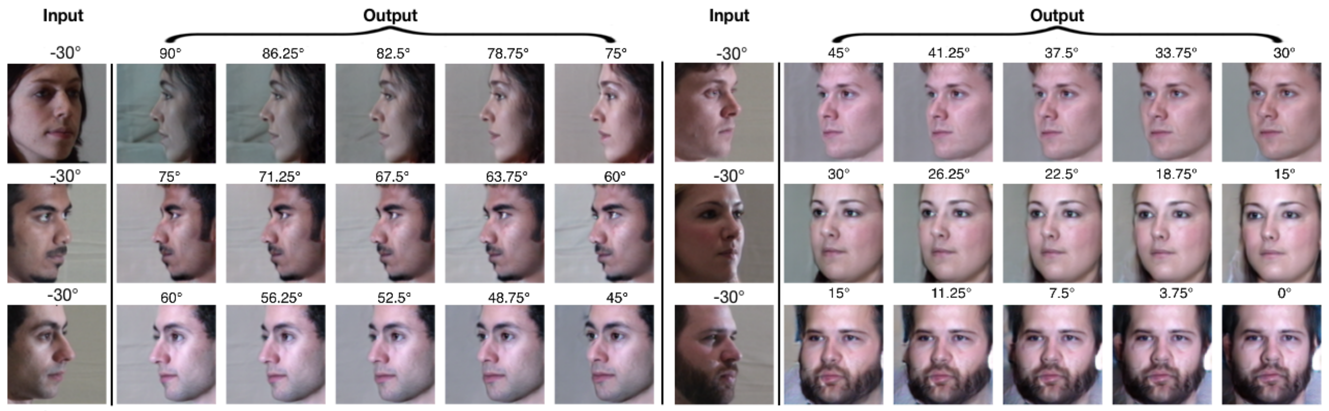

Figure 3: Face rotation results on Multi-PIE. Each input is rotated to the specified yaw angle which is indicated by the degree above the output.

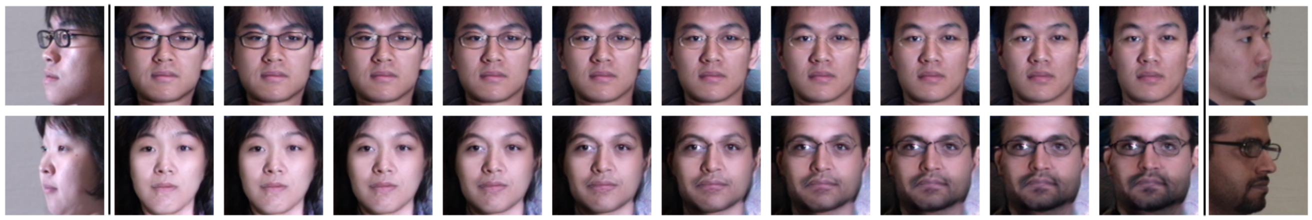

Figure 4: Synthesizing new samples through interpolating identity representations. For each row, the samples in the middle are synthesized by the interpolation of the identity representations of the far left and the far right faces.

codes for 15° and 0° to get the one for 7.5°. It will be hard for human observers to find the evidence of forgery on our results. Samples synthesized by interpolating identity representations [Tran *et al.*, 2017] are reported in Fig. 4. Given two images of different subjects, we extract identity representations from the bottleneck layer of G_E and then generate new representations through interpolation. Fed by those new representation, G_E will synthesize new images with "fused" identities. We can see that the semantic changes in those images are very smooth and the visual realism is also very desirable. Frontalization results on IJB-A are shown in Fig. 5. We compare with results produced by DR-GAN [Tran *et al.*, 2017]. The background of the input and the produced images looks different. We argue that for face frontalization in such challenging unconstrained environments, the concentration should be put on the facial parts. Despite that yaw angles are very large, our model still produces plausible results. DR-GAN produces comparable results, but the identities of some produced samples look very different from the original inputs.

Pose-invariant Face Recognition. To evaluate the capacity of identity preservation, we first use our model to frontalize profile face images in Multi-PIE and then evaluate face recognition performances through those produced images. We employ LightCNN [He *et al.*, 2017] as our feature extractor. We make comparisons with DR-GAN, TP-GAN [Huang *et al.*, 2017] and FF-GAN [Yin *et al.*, 2017]. The results are reported in Table 1. Note that TP-GAN and FF-GAN can only produce frontal view images. It can be observed that our results outperform them, especially for extreme poses, which indicates that our LB-GAN is able to preserve better identity information.

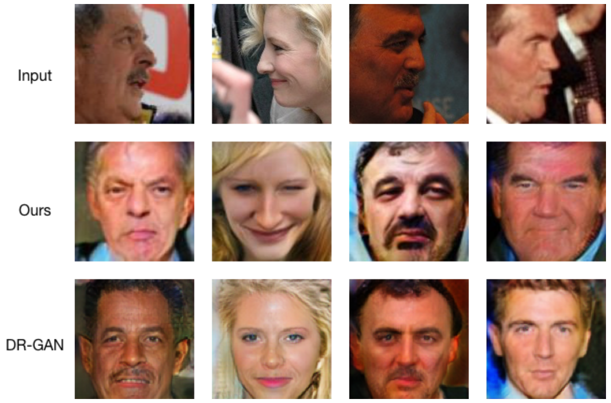

Figure 5: Synthesis results on the IJB-A dataset.

Head Pose Estimation. To test whether our model is able to give correct responses to the remote code, we make head pose estimations on Multi-PIE. A **t**hird-party **h**ead **p**ose **e**stimator (THPE) [1] is employed. We simply call the high-

[1] https://github.com/guozhongluo/

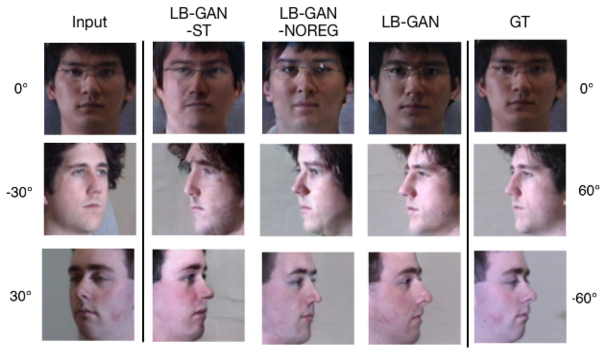

Figure 6: Qualitative comparisons on synthesis results between LB-GAN and its variants. The degrees on the left side are the yaw angles of the inputs, and those on the right side are set by the remote codes.

Table 3: Identification rate (%) comparison of model variations of our LB-GAN on Multi-PIE.

Method	±15°	±30°	±45°	±60°	±75°	±90°
LB-GAN	**99.1**	**98.9**	**96.7**	**91.0**	**80.3**	**65.4**
LB-GAN-ST	93.7	92.1	85.3	83.2	72.0	59.9
LB-GAN-NOREG	96.8	93.9	91.2	87.6	77.6	63.5

level interface to train the model and then get the predicted yaw angles. The output of THPE is continuous angle value. Note that THPE is trained on the 300W dataset [Sagonas *et al.*, 2013] which consists only face images with yaw angles within ±30°. So only the images whose yaw angles are within this range are tested. The mean pose estimation errors are reported in Tabel 2. Those images with yaw angles of ±7.5° and ±22.5° are produced through interpolation. We can see that the mean errors made by THPE on the genuine and the synthesized data across all poses are very similar, which indicates that LB-GAN has the ability to control the pose of the output. The error of the images produced by interpolation is higher but still within an acceptable range.

4.3 Ablation Study

In this section, we demonstrate the effectiveness of our proposed two-stage training method and regularization items through an ablation study. Both qualitative and quantitative results are compared to make a comprehensive understanding. Specifically, we investigate the following model variations:

- LB-GAN-ST: All the components of the networks are trained jointly in a single stage through the conventional manner [Goodfellow *et al.*, 2014].
- LB-GAN-NOREG: The network is trained in the same way as LB-GAN. L_{csc} is removed. $L2'$ loss is replaced by $L2$ loss.

A visual comparison is shown in Fig. 6. The face recognition performances on Multi-PIE are reported in Table 3. It can be observed that LB-GAN produces the most visually appealing results as well as achieve the best verification performance. The inferior performance of LB-GAN-ST indicates that two-stage training method is very important for our model. The effectiveness of regularization items is validated by the comparison between LB-GAN-NOREG and LB-GAN. As shown by the top row of images, the model tends to make superfluous manipulations and changes the contour of the input face obviously without those regularization items. Those observations prove that the regularization items can guide our model to put more attention on optimizing the facial part of its output.

5 Conclusion

This paper has proposed LB-GAN for multi-view face image synthesis by decomposing the synthesis process into two subtasks. Input face images are first transformed into a frontal view by the face normalizer and then rotated to a specified angle by the face editor. A novel two-stage training method has also been accordingly proposed to help accomplish the two subtasks smoothly. To further improve the performance, conditional self-cycle loss and improved $L2$ loss have been integrated into LB-GAN. Experimental results have shown that our method is able to alter the pose of an input face image and keep the visual appearance photo-realistic simultaneously. Besides, our method obtains state-of-the-art face recognition results on publicly available datasets. In the future, we will investigate on the method to control more facial attributes, e.g., expression, race and gender.

References

[Blanz and Vetter, 1999] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[Dovgard and Basri, 2004] Roman Dovgard and Ronen Basri. Statistical symmetric shape from shading for 3d structure recovery of faces. In *ECCV*, pages 99–113, 2004.

[Ferrari *et al.*, 2016] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Effective 3d based frontalization for unconstrained face recognition. In *ICPR*, pages 1047–1052, 2016.

[Ghodrati *et al.*, 2016] Amir Ghodrati, Xu Jia, Marco Pedersoli, and Tinne Tuytelaars. Towards automatic image editing: Learning to see another you. In *BMVC*, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *IVC*, 28(5):807–813, 2010.

[Hassner *et al.*, 2015] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *CVPR*, pages 4295–4304, 2015.

[Hassner, 2013] Tal Hassner. Viewing real-world faces in 3d. In *ICCV*, pages 3607–3614, 2013.

`head-pose-estimation-and-face-landmark`

[He *et al.*, 2017] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, pages 2000–2006, 2017.

[Huang *et al.*, 2017] 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.

[Kan *et al.*, 2014] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, pages 1883–1890, 2014.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Klare *et al.*, 2015] Brendan F. Klare, Anil K. Jain, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Mark Burge. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. pages 1931–1939, 2015.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[Sagonas *et al.*, 2013] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 4, page 7, 2017.

[Yang *et al.*, 2015] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, pages 1099–1107, 2015.

[Yi *et al.*, 2014] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[Yim *et al.*, 2015] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, pages 676–684, 2015.

[Yin *et al.*, 2017] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[Zhang *et al.*, 2013] Yizhe Zhang, Ming Shao, Edward K Wong, and Yun Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *ICCV*, pages 2416–2423. IEEE, 2013.

[Zhao *et al.*, 2017] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, pages 65–75, 2017.

[Zhu *et al.*, 2013] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *ICCV*, pages 113–120, 2013.

[Zhu *et al.*, 2014] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, pages 217–225, 2014.

[Zhu *et al.*, 2015] 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*, pages 787–796, 2015.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.