Pose-Guided Photorealistic Face Rotation

Yibo Hu[1,2], Xiang Wu[1], Bing Yu[3], Ran He[1,2]*, Zhenan Sun[1,2]

[1]CRIPAC & NLPR & CEBSIT, CASIA [2]University of Chinese Academy of Sciences
[3]Noah's Ark Laboratory, Huawei Technologies Co., Ltd.
{yibo.hu, xiang.wu}@cripac.ia.ac.cn, yubing5@huawei.com, {rhe, znsun}@nlpr.ia.ac.cn

Abstract

Face rotation provides an effective and cheap way for data augmentation and representation learning of face recognition. It is a challenging generative learning problem due to the large pose discrepancy between two face images. This work focuses on flexible face rotation of arbitrary head poses, including extreme profile views. We propose a novel Couple-Agent Pose-Guided Generative Adversarial Network (CAPG-GAN) to generate both neutral and profile head pose face images. The head pose information is encoded by facial landmark heatmaps. It not only forms a mask image to guide the generator in learning process but also provides a flexible controllable condition during inference. A couple-agent discriminator is introduced to reinforce on the realism of synthetic arbitrary view faces. Besides the generator and conditional adversarial loss, CAPG-GAN further employs identity preserving loss and total variation regularization to preserve identity information and refine local textures respectively. Quantitative and qualitative experimental results on the Multi-PIE and LFW databases consistently show the superiority of our face rotation method over the state-of-the-art.

1. Introduction

Benefiting from the convolutional neural networks trained on large-scale face databases [2, 10], the performance of face recognition systems has been significantly improved in recent years. However, pose variations still pose a great challenge to face recognition in real-world scenarios. The existing methods that address this pose problem can be generally categorized into two classes. The first one aims to obtain pose-invariant embeddings [2, 22, 28]. The other aims to normalize face images [14, 35] to frontal views, which can be directly used by general face recognition methods without retraining the recognition models.

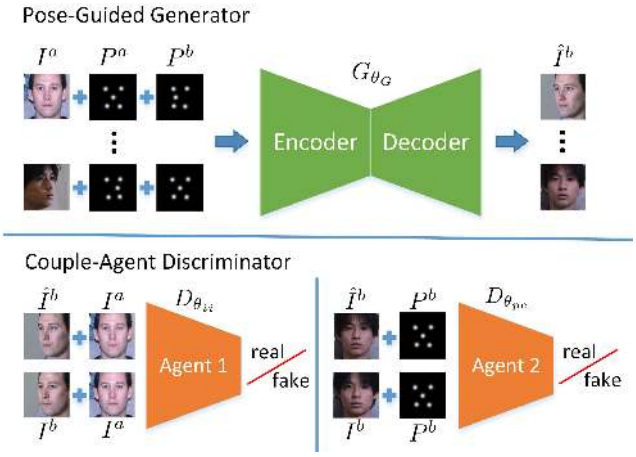


Figure 1. The framework of CAPG-GAN. The generator incorporates pose information by landmark heatmaps in the learning process. The couple-agent discriminator distinguishes generated pairs from ground-truth pairs for photorealistic synthesization.

For the first class, metric learning [25] is a common way to achieve pose-invariant embeddings. Moreover, multi-view [19, 30] or multiple pose-specific [21] methods are also used to obtain pose invariance. However, due to the imbalanced distributions that characterize a long tail distribution of large pose faces, it is often difficult to achieve ideal pose-invariance across large pose variations.

The second class is often known as face rotation, which resorts to computer graphics or deep learning to rotate profile faces to frontal views. Recently, deep learning based methods have shown impressive capability on face rotation [17, 27, 39]. Early efforts [34] adopt the mean square loss to learn a deep regression model from paired training data. Later on, multiple methods have been proposed with novel network architecture [33, 40, 41] or learning objectives [14, 28, 35]. As a representative method, TP-GAN [14] adopts a two-pathway architecture along with sophisticated loss functions to learn photorealistic frontal view synthe-

*corresponding author

sis. However, we argue that TP-GAN has two limitations. Firstly, it employs a complex architecture to preserve global and local texture information. It contains a global network and four local patch networks whose training and inference are very time-consuming. Secondly, its architecture and loss designment are specific for face frontalization and are not applicable for arbitrary pose synthesis. In real world scenarios, arbitrary pose synthesis is more appealing since it has more potential applications, such as face edition and data-augmentation for face representation learning.

To address the above issues, this paper proposes a novel **C**ouple-**A**gent **P**ose-**G**uided **G**enerative **A**dversarial **N**etwork (CAPG-GAN) for fast face rotation of arbitrary poses. Specially, we introduce landmark heatmaps of both the input and target faces to the generator network to incorporate pose information in the learning process. Landmark heatmaps of the target face provide guidance for arbitrary pose synthesis, while that of the input face serves to capture the local-aware information as the local pathway in TP-GAN. By replacing the local pathway with one more input (landmark heatmaps), training and inference speed can be substantially accelerated. In addition, we propose a couple-agent discriminator to harness target pose information. The couple agents reinforce on discriminating the realism of synthetic arbitrary view faces from the pose-guided generator. Moreover, an identity preserving constraint is implemented by Light CNN [29]. Under these configurations, the pose-guided generator needs to play against the couple-agent discriminator and is reinforced by an identity preserving network to synthesize arbitrary pose faces that are useful for face recognition.

To summarize, the main contributions are as follows:

- A Couple-Agent Pose-Guided Generative Adversarial Network (CAPG-GAN) is proposed for face rotation from a single image in 2D space, which can synthesize arbitrary view images. It can not only frontalize a face for face recognition, but also rotate faces to an arbitrary pose.
- Landmark heatmaps are newly used as a controllable signal in pose-guided generator to synthesize face images. Compared to the previous methods directly based on pose degree or 3D information, the heatmap provides a flexible and efficient way for both learning and inference.
- The proposed couple-agent discriminator efficiently combines prior domain knowledge of pose and local structure of face to reinforce the realism of synthetic arbitrary view faces.
- Our CAPG-GAN demonstrates the possibility to synthesize photorealistic and identity-preserving faces with arbitrary poses and achieves state-of-the-art face recognition performance under large pose variations.

The rest of this paper is organized as follows. We briefly review some related works in Section 2. In Section 3, we present the details of our Couple-Agent Pose-Guided Generative Adversarial Network (CAPG-GAN). The experimental results and algorithmic analyses are shown in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

2.1. Generative Adversarial Network

Introduced by Goodfellow *et al*. [8], Generative Adversarial Network (GAN) plays a min-max game to improve both discriminator and generator. With the constraints of the min-max game, GAN can encourage the generated images to be close to the true image manifolds. Recently, deep convolutional generative adversarial network (DCGAN) [23] has demonstrated the superior performance of image generation. Info-GAN [3] applies information regularization to optimization. Furthermore, Wasserstein GAN [1] improves the learning stability of GAN and provides solutions of debugging and hyperparameter searching for GAN. These successful theoretical analyses of GAN show the effectiveness and possibility of photorealistic face image generation and synthesis.

2.2. Face Frontalization

Face frontalization is an extremely challenging synthesis problem due to its ill-posed nature. Traditional methods addressing this problem can be divided into three categories: 3D/2D local texture warping [12, 39], statistic modeling [24] and deep learning based methods [4, 14, 18, 33, 34, 35]. Hassner *et al*. [12] employ a single unmodified 3D reference surface to produce frontal view. A joint frontal view reconstruction and landmark localization are optimized by the minimization of the nuclear norm in [24].

With the development of deep learning, Kan *et al*. [18] propose SPAE for face frontalization via auto-encoders. Yim *et al*. [34] introduce a multi-task learning for frontal view synthesis. Yang *et al*. [33] employ a recurrent transformation unit to synthesize discrete 3D views. Moreover, Cole *et al*. [4] decompose faces into a sparse set of landmarks and aligned texture maps by a network, and then combine them by a differentiable image warping operation.

Benefiting from GAN, Huang *et al*. [14] propose TP-GAN, which processes profile view faces through global and local networks separately. Some domain knowledge such as symmetry and identity information of face is used to make the synthesized faces photorealistic. Towards large-pose face frontalization in the wild, FF-GAN [35] is proposed to incorporate 3D face model into GAN. In this way, the 3DMM conditioned GAN can retain the visual quality under occlusions during frontalization.

2.3. Recognition via Generation

Face frontalization provides a practical way for recognition via generation. Recognition can be directly performed on the synthesized frontal face image without retraining the recognition models. With the help of generative models, recognition via generation has drawn much attention recently. Benefiting from dual learning [31], introducing generative procedure into discriminative tasks is an effective way to improve the performance. Tran *et al*. [28] propose DR-GAN to learn a generative and discriminative representation in addition to image generation. The representation is disentangled from pose variations through pose embedding in the generator and pose estimation in the discriminator. DR-GAN achieves significant improvements in pose-invariant face recognition on various face databases. DA-GAN [37] is proposed to synthesize profile face images by adversarial training. It uses a 3D face model to simulate profile face and then recovers the lost information inherent from 3D model space to 2D image space. The qualitative and quantitative experiments on IJB-A [20] show the effectiveness of recognition via generation framework. Moreover, Song *et al*. [26] introduce adversarial learning into a cross-spectral face hallucination, facilitating heterogeneous face recognition via generation.

3. Approach

Face rotation aims to synthesize an arbitrary pose face image I^b from a given pose face image I^a. Our goal is to learn such a synthesizer that can infer the corresponding view images. Particularly, we model the synthesizer as a CNN with u-net [15] architecture. In order to make the synthesized images photorealistic and preserve their identity information, we introduce a couple-agent discriminator with an identity preserving constraint to incorporate prior knowledge from data distribution and domain knowledge of faces. Thus the synthesizer can exactly recover and reconstruct the lost information inherent in face rotation.

The overall framework of our proposed Couple-Agent Pose-Guided GAN (CAPG-GAN) is depicted in Fig. 1, which mainly contains two parts: a pose-guided generator and a couple-agent discriminator. Both of them are supervised by a weighted sum of several losses in an end-to-end manner that jointly force the synthesized face images to be photorealistic and identity preserving. In the rest of this section, we will first introduce the architectures of the generator and the discriminator. Then, we details all the supervised loss functions.

3.1. Network Architecture

3.1.1 Pose-Guided Generator

In order to rotate face images to arbitrary poses, our generator should incorporate the pose information. Different from the recent work [28] that encodes pose information to an one-hot vector and the work [37] that employs a 3DMM to involve pose information, our generator can adaptively learn such information from pose embeddings, i.e. the rotation angles are learned by our model without knowing in advance. The pose embeddings are obtained by an off-the-shelf facial landmark detector [36]. We first apply the detector to estimate the coordinates of 5 facial landmarks. Then we encode them as 5 heatmaps, named pose embeddings. Each heatmap is filled by a Gaussian distribution where the mean locates at the corresponding coordinate and the standard deviation is set to 2.

As shown in Fig. 1, we concatenate the source image I^a, the source pose embeddings P^a and the target pose embeddings P^b as the inputs of our generator G_{θ_G} and synthesize the target view image $\hat{I}^b$. That is,

$$\hat{I}^b = G_{\theta_G}\left(I^a, P^a, P^b\right) \tag{1}$$

where θ_G is the parameter of our generator. Guided by the pose embeddings, our generator can synthesize face images under various poses.

Inspired by the recent success of u-net architecture in image-to-image translation [15, 38], our G_{θ_G} consists of a down-sampling encoder and an up-sampling decoder with skip connections for multi-scale feature fusion. Such architecture comprehensively retains contextual and textural information, which is crucial for removing artifacts and padding textures. The detailed architecture of G_{θ_G} is listed in Supplementary Materials.

3.1.2 Couple-Agent Discriminator

To exploit prior domain knowledge from data distribution (pose and facial structure information), we extend the standard discriminator in vanilla GAN [8] with a couple-agent structure, as shown in Fig. 1.

Agent 1 is a conditional discriminator $D_{\theta_{ii}}$ implemented by a CNN framework, where θ_{ii} is the parameter. It takes the source image I^a as the condition and pairs the output of generator $\hat{I}^b$ (or the target natural image I^b) with the condition image I^a as its input. Our $D_{\theta_{ii}}$ can not only distinguish synthesized and natural face images, but also learn the distinction of rotated poses. It maps the input pairs to a probability map instead of one scalar value, as presented in [14]. Now, each position in the probability map corresponds to a local region instead of the whole face. Thus $D_{\theta_{ii}}$ specifically focuses on each semantic region.

Agent 2 is also a conditional discriminator $D_{\theta_{pe}}$, having the same architecture with agent 1, where θ_{pe} is the parameter. Specially, it employs the target pose embeddings P^b as the condition and pairs the output of generator $\hat{I}^b$ (or the target natural image I^b) with the condition pose embeddings

P^b as its input. This pairwise input encourages $D_{\theta_{pe}}$ to discriminate the diversity of facial structure and capture the local-aware information. The detailed architecture is listed in Supplementary Materials.

3.2. Training Losses

Our CAPG-GAN is supervised by a weighted sum of several losses in an end-to-end manner, including multi-scale pixel-wise loss, conditional adversarial loss, identity preserving loss and total variation regularization.

3.2.1 Multi-Scale Pixel-Wise Loss

We employ multi-scale pixel-wise L1 loss on the synthesized face image $\hat{I}^b$ to constrain the content consistency,

$$L_{pix} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{W_s H_s C}\sum_{w,h,c=1}^{W_s,H_s,C}\left|\hat{I}^b_{s,w,h,c} - I^b_{s,w,h,c}\right| \quad (2)$$

where S denotes the number of scales. W_s and H_s represent the width and height of each image scale, respectively. C is the number of image channel. We choose the last three scales (32×32, 64×64 and 128×128) of the synthesized images produced by the decoder in our generator to calculate this loss. The multi-scale pixel-wise loss may smooth the synthesized images, but it is crucial for speeding up optimization and reconstructing the global information.

3.2.2 Conditional Adversarial Loss

To incorporate prior domain knowledge from data distribution (pose and facial structure information) and remove smoothness caused by multi-scale pixel-wise loss, we introduce conditional adversarial loss to our CAPG-GAN. The conditional adversarial loss of $D_{\theta_{ii}}$ for distinguishing synthesized image pairs $\{\hat{I}^b, I^a\}$ from real image pairs $\{I^b, I^a\}$ is formulated as follows:

$$\begin{aligned} L^{ii}_{adv} = {} & E_{I^b\sim P(I^b)}\left[\log D_{\theta_{ii}}\left(I^b, I^a\right)\right] + \\ & E_{\hat{I}^b\sim P(\hat{I}^b)}\left[\log\left(1 - D_{\theta_{ii}}\left(\hat{I}^b, I^a\right)\right)\right] \end{aligned} \quad (3)$$

The conditional adversarial loss of $D_{\theta_{pe}}$ for distinguishing pairs $\{\hat{I}^b, P^b\}$ from pairs $\{I^b, P^b\}$ takes the form,

$$\begin{aligned} L^{pe}_{adv} = {} & E_{I^b\sim P(I^b)}\left[\log D_{\theta_{pe}}\left(I^b, P^b\right)\right] + \\ & E_{\hat{I}^b\sim P(\hat{I}^b)}\left[\log\left(1 - D_{\theta_{pe}}\left(\hat{I}^b, P^b\right)\right)\right] \end{aligned} \quad (4)$$

where L^{ii}_{adv} and L^{pe}_{adv} aim to preserve pose information and reconstruct local structure information respectively. Both of them contribute to visually pleasing synthesized images.

3.2.3 Identity Preserving Loss

Supervised by the above two losses, our model can produce photorealistic face images, but these images are weak for recognition due to lack of identification information. To integrate domain knowledge of identities, we exploit an identity preserving network D_{ip} to preserve the identity discrimination of the synthesized face images, which is derived from [28, 37, 14]. D_{ip} is a feature extractor that can force the features extracted from synthesized images $\hat{I}^b$ to be as close to the features extracted from target images I^b as possible. It makes the same subject form a compact cluster with small intra-class distances and variances in embedding space. We choose a pre-trained Light CNN [29] as D_{ip} and fix the parameters during training procedure. Specifically, we define the identity preserving loss on the output of the last pooling layer and the fully connected layer of D_{ip}:

$$L_{ip} = \left\|D^p_{ip}(\hat{I}^b) - D^p_{ip}(I^b)\right\|^2_F + \left\|D^{fc}_{ip}(\hat{I}^b) - D^{fc}_{ip}(I^b)\right\|^2_2 \quad (5)$$

where $D^p_{ip}(\cdot)$ and $D^{fc}_{ip}(\cdot)$ denote the output of the last pooling layer and the fully connected layer, respectively. $\|\cdot\|_2$ means the vector 2-norm while $\|\cdot\|_F$ represents the matrix F-norm.

3.2.4 Total Variation Regularization

Usually, the images synthesized by GAN model have many unfavorable artifacts [23], which deteriorate the visualization and the recognition performance. We impose a total variation regularization term [16] on the final synthesized images to alleviate this issue:

$$L_{tv} = \sum_{c=1}^{C}\sum_{w,h=1}^{W,H}\left|\hat{I}^b_{w+1,h,c} - \hat{I}^b_{w,h,c}\right| + \left|\hat{I}^b_{w,h+1,c} - \hat{I}^b_{w,h,c}\right| \quad (6)$$

where W and H represent the width and height of the final synthesized images.

3.2.5 Overall Loss

The total supervised loss is a weighted sum of the above losses. The pose-guided generator G_{θ_G}, couple-agent $D_{\theta_{ii}}$ and $D_{\theta_{pe}}$ are trained alternatively to optimize the following min-max problem:

$$\min_{\theta_G}\max_{\theta_{ii},\theta_{pe}} L = \lambda_1 L_{pix} + \lambda_2 L^{ii}_{adv} + \lambda_3 L^{pe}_{adv} + \lambda_4 L_{ip} + \lambda_5 L_{tv} \quad (7)$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are the trade-off parameters. Solving this min-max problem makes our generator synthesize various view face images guided by pose embeddings in photorealistic and identity preserving manners.

4. Experiments

CAPG-GAN provides a flexible way to rotate an input face to any pose controlled by facial landmarks. It can be used for both photorealistic face synthesis and pose invariant representation learning. For the former, we show qualitative results of both face frontalization and profile face generation. For the latter, we quantitatively evaluate face recognition performance based on the synthesized face images under both the controlled and in-the-wild settings. We extensively compare CAPG-GAN with state-of-the-art methods on the Labeled-Faces-in-the-Wild (LFW) database [13] and the Multi-PIE database [9], which have been widely used for face synthesis or recognition [32][35][14]. In the following subsections, we begin with an introduction of settings and databases. Then we demonstrate the merits of our CAPG-GAN on qualitative synthesis results and quantitive verification results over state-of-the-art methods. Lastly, we conduct ablation study to demonstrate the benefits gained from each part of CAPG-GAN.

4.1. Settings and Databases

Databases. The LFW database [13] contains 13,233 images of 5,749 people. It has been widely used to evaluate synthesis or verification performance of various methods under unconstrained environments. Since the face images in LFW are collected from the web and contain various pose, expression and illumination variations, it is extremely challenging to synthesize a photorealistic frontal face. For the verification protocol [13], face images are divided in 10 folds that contain different identities and 600 face pairs. We evaluate face verification performance on the synthesized images and compare CAPG-GAN with previous face synthesized or rotation methods.

The Multi-PIE database [9] is the largest database for evaluating face synthesis and recognition in the controlled setting. There are four sessions in this database. The face images from 337 identities have 20 illumination levels and 15 poses ranging from -90^o to 90^o. In this paper, the face images with neutral expression under 20 illuminations and 13 poses within $\pm 90^o$ are used. The two poses from the two additional cameras (08_1 and 19_1) located above the subject are not considered. We follow the testing protocols in [32][34][14] and unitize two settings to evaluate different methods.

In the first setting [32], we only use the images from the session 1, which contains faces of 250 subjects. The training set is composed of all the images (13 poses and 20 illumination levels) of the first 150 identities, i.e., $150 \times 13 \times 20 = 39000$ images in total. For testing, one gallery image with frontal view and normal illumination is used for each of the remaining 100 subjects. The numbers of the probe and gallery sets are 24,000 and 100 respectively. The second setting, followed by the protocol from [34], includes neutral expression images from all four sessions. We use the first 200 subjects and the remaining 137 subjects for training and testing respectively. Each testing identity has one gallery image from his first appearance. Hence, there are 161,460, 72,000, 137 images in the training, probe and gallery sets respectively. Note that, for the two settings, there are no overlap subjects between the training and testing sets.

Implementation Details. To train CAPG-GAN, pairs of images $\{I^a, I^b\}$ from multiple poses and identities are required. Enumerating all the corresponding pairs is infeasible and time consuming during training. To ease it, we simplify the training procedure as a bidirectional face rotation task: one is from profile view with arbitrary poses to frontal view, the other one is from frontal view to profile view with arbitrary poses. That means the corresponding pairs $\{I^a, I^b\}$ consist of one frontal face image and one profile face image, but which one comes from frontal view is stochastic during training. RGB images of a size 128×128 are used in all experiments for both real and synthesized images. Our identity preserving network is trained on MS-Celeb-1M [11] and fine-tuned on the real training images of Multi-PIE. Our model is implemented with Pytorch. In all our experiments, we empirically set $\lambda_1 = 10$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.02$, $\lambda_5 = 1e-4$. The learning rate is initialized by 0.0002 and linearly decayed after each epoch until reaching 0.

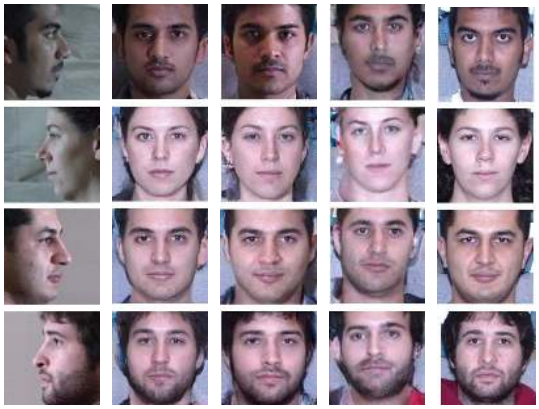
4.2. Face Synthesis

In this subsection, we systematically compare the synthesis results of CAPG-GAN against state-of-the-art face synthesis methods. CAPG-GAN is trained on the training set of the Setting 2 from the Multi-PIE database. The synthesis results are verified on the testing set of the Setting 2 and the LFW database. Hence, there are no overlap subjects between the training and testing sets. Since most previous frontal view synthesis methods are dedicated to address face synthesis problem within a small pose range of $\pm 60°$, we firstly show the synthesis results of different methods under the pose of $45°$ and $30°$ in Fig. 2 and then present large pose situation with $75°$ and $90°$ in Fig. 3. Particularly, the synthesized results of TP-GAN and our CAPG-GAN are obviously better than other methods in terms of global structure and local texture. The results of our CAPG-GAN are comparable to or better than those of TP-GAN. Compared to TP-GAN that can only synthesize a frontal face, our CAPG-GAN is more flexible because it resorts to landmarks as a condition to rotate a face image. Moreover, the network structure of CAPG-GAN is simpler than that of TP-GAN.

Note that, different from previous methods, CAPG-GAN can rotate an input face to any pose controlled by facial landmarks. Hence, CAPG-GAN can synthesize both frontal and profile faces. Fig. 5 further plots the synthesized frontal

(a) Profile (b) Ours (c) [14] (d) [28] (e) [34] (f) [7] (g) [39] (h) [12] (i) Frontal

Figure 2. Synthesis results of different methods under the pose of 45° (first two rows) and 30° (last row).



(a) Profile (b) Ours (c) [14] (d) [35] (e) Frontal

Figure 3. Synthesis results of different methods under the pose of 75° (first two rows) and 90° (last two rows).

(a) LFW (b) Ours (c) [14] (d) [39]

Figure 4. Visual comparison of face synthesis on the LFW database. Our method is trained on Mulit-PIE and tested on LFW.

and profile results by CAPG-GAN under poses 90°, 75°, 60°, 45°, 30° and 15°. The first and third images in each column are ground truth, and the second and forth images are synthesized results. Moreover, the synthesis results of different target pose embeddings are shown in Fig. 6. We observe that CAPG-GAN can not only well preserve the overall facial structure but also recover the unseen ears and cheeks in an identity consistent way. Both synthesized frontal and profile faces have good qualities. These good results may benefit from the heatmaps of facial landmarks. The heatmaps potentially guide CAPG-GAN to synthesize some particular facial areas and reconstruct local structure that have large variations during face rotation. These results also suggest that given enough training data and a proper network structure, it is feasible to synthesize a photorealistic frontal face image from a large pose.

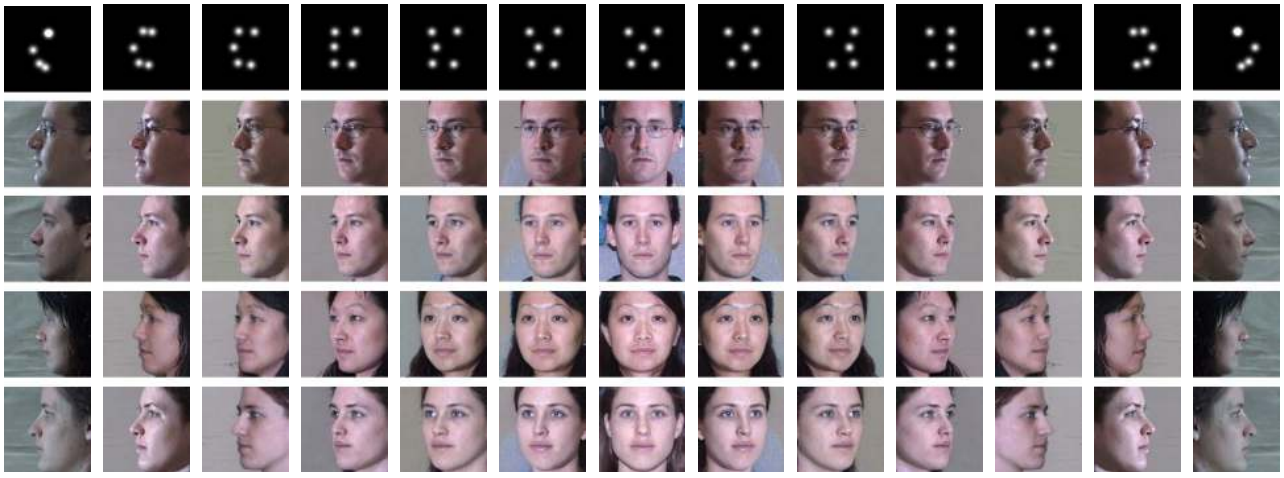
To further verify the synthesis ability of CAPG-GAN in unconstrained environments, we perform visual comparison of face synthesis on the LFW database. Fig. 4 shows some synthesis results of different methods. Note that our method is only trained on Mulit-PIE and tested on LFW. As expected, CAPG-GAN can also obtain good visual results on LFW. Our visual results are obviously better than [39] and comparable to [14].

4.3. Identity Preserving Property

In face synthesis, recognition accuracy is commonly used to quantitatively evaluate the identity preserving ability of different methods. Face synthesis is an ill-posed problem. An original face image may not contain all pixel-level information of one subject. A face synthesis method can generate the lost pixel-level information but may also lose some identity information. If the synthesized images can result in better recognition accuracy, more identity information is preserved during synthesis process. Hence, in this subsection, we quantitatively compare the proposed method with other methods in terms of recognition accuracy.

Table 1 shows the Rank-1 accuracies of different methods under the Setting 1 of Multi-PIE. The accuracies of all methods drop as pose degree increases. This is because more facial appearance information is lost and synthesis task becomes more difficult when pose degree increases. The results of Light CNN are used as the baseline of our method. We compare CAPG-GAN with CPF [34], Hass-

Figure 5. Synthesized frontal and profile results by CAPG-GAN under different poses. From top left to bottom right, the poses are 90°, 75°, 60°, 45°, 30° and 15°. The first and third images in each column are ground truth, and the second and forth images are synthesized.



(a) $+90^\circ$ (b) $+75^\circ$ (c) $+60^\circ$ (d) $+45^\circ$ (e) $+30^\circ$ (f) $+15^\circ$ (g) 0° (h) -15° (i) -30° (j) -45° (k) -60° (l) -75° (m) -90°

Figure 6. Synthesis results of different target pose embeddings.

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

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
CPF[34]	-	-	-	71.65	81.05	89.45
Hassner *et al*. [12]	-	-	44.81	74.68	89.59	96.78
HPN[5]	29.82	47.57	61.24	72.77	78.26	84.23
FIP_40[40]	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN Forest[32]	47.26	60.66	74.38	89.02	94.05	96.97
TP-GAN[14]	64.03	84.10	92.93	**98.58**	**99.85**	99.78
Light CNN[29]	9.00	32.35	73.30	97.45	99.80	99.78
CAPG-GAN	**77.10**	**87.40**	**93.74**	98.28	99.37	**99.95**

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

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
FIP+LDA[40]	-	-	45.9	64.1	80.7	90.7
MVP+LDA[41]	-	-	60.1	72.9	83.7	92.8
CPF[34]	-	-	61.9	79.9	88.5	95.0
DR-GAN[28]	-	-	83.2	86.2	90.1	94.0
FF-GAN[35]	61.2	77.2	85.2	89.7	92.5	94.6
TP-GAN[14]	64.64	77.43	87.72	95.38	98.06	98.68
Light CNN[29]	5.51	24.18	62.09	92.13	97.38	98.59
CAPG-GAN	**66.05**	**83.05**	**90.63**	**97.33**	**99.56**	**99.82**

ner *et al*. [12], HPN [5], FIP_40 [40], c-CNN Forest [32] and TP-GAN [14], and observe that CAPG-GAN not only significantly outperforms its competitors under challenging $\pm 90^o$ but also achieves the best or comparable performance across other angles. It seems that the larger the head pose, the greater the improvement is obtained by CAPG-GAN. These quantitative results demonstrate that CAPG-GAN lose less identity information during face synthesis.

Table 2 further tabulates the results of different methods under the Setting 2 of Multi-PIE. The Setting 2 is challenging than the Setting 1. This is because there are 72,000 and 137 images in probe and gallery sets respectively. The probe set contains large appearance variations whereas the gallery set only contains one image per subject. Once more, the results of Light CNN are used as the baseline of our method. We observe that the methods can be ordered in ascend-

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

Method	ACC(%)	AUC(%)
Ferrari *et al.* [6]	-	94.29
LFW-3D[12]	93.62	88.36
LFW-HPEN[39]	96.25	99.39
FF-GAN[35]	96.42	99.45
CAPG-GAN	**99.37**	**99.90**

Table 4. Model comparison: Rank-1 recognition rates (%) under Setting 2.

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
w/o L_{ip}	40.77	46.13	53.25	64.74	76.16	86.12
w/o L_{tv}	61.33	78.98	87.68	95.58	99.03	99.74
w/o L_{adv}	46.83	56.90	67.68	85.68	96.26	99.50
w/o L_{adv}^{ii}	54.68	66.09	75.90	89.38	97.79	99.73
w/o L_{adv}^{ip}	57.78	71.17	82.05	92.50	97.57	99.63
CAPG-GAN	**66.05**	**83.05**	**90.63**	**97.33**	**99.56**	**99.82**

ing Rank-1 accuracy as FIP+LDA [40], MVP+LDA [41], CPF [34], DR-GAN [28], FF-GAN [35], TP-GAN [14] and CAPG-GAN. CAPG-GAN is significantly better than other methods. Particularly, the same as TP-GAN, CAPG-GAN also directly uses the synthesized images to perform recognition and follows the way of recognition via generation.

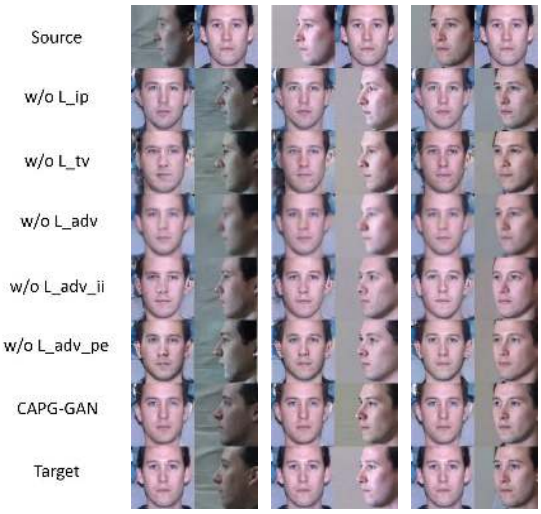
Besides, Table 3 shows the accuracies of different methods under in-the-wild setting. As expected, CAPG-GAN achieves 99.37% on accuracy and 99.90% on AUC, which are also comparable with other state-of-the-art methods. Although CAPG-GAN is not trained on the LFW database, its synthesized face images can also further improve recognition accuracy, suggesting the identity preserving character of CAPG-GAN in the wild.

4.4. Ablation Study

In this subsection, five loss function combinations in CAPG-GAN are studied to give an insight into their respective roles. We report both qualitative visualization results and quantitative recognition results for a comprehensive comparison.

Fig. 7 plots visual comparisons between CAPG-GAN and its five incomplete variants. We observe that CAPG-GAN is visually better than its variants across all poses. Without the L_{ip} loss, the local textures around eyes on the generated profile faces are few and the facial contours are distorted. Without the L_{tv} loss, there are artifacts on the generated faces. Without the L_{adv} loss (both L_{adv}^{ii} and L_{adv}^{pe}), the generated faces tend to be blur, suggesting the usage of adversarial learning. Without the L_{adv}^{ii} loss, the rotation around nose on the generated frontal faces is incomplete, indicating the agent 1 discriminator dominates rotation validity. Without the L_{adv}^{pe} loss, the structures of eyes and mouth have deformations, revealing the agent 2 discriminator enforces facial structures. These visual results demonstrate that each component in CAPG-GAN is essential to obtain a photorealistic face image.

Table 4 further tabulates the quantitative results of different variants of CAPG-GAN. The recognition accuracy is evaluated on the synthetic images generated from each variant. We observe that the accuracy will decrease if one loss function is not used. Particularly, the accuracy drops significantly for all poses if the L_{ip} loss is not adapted. These results suggest that each component in CAPG-GAN is essential for identity preserving during synthesis.



(a) (b) 90° (c) 60° (d) 30°

Figure 7. Model comparison: synthesis results of CAPG-GAN and its variants.

5. Conclusion

This paper has proposed a CAPG-GAN method to synthesize a face under various head poses, including extreme profile views. CAPG-GAN can synthesize both neutral and profile head pose face images. It has encoded the head pose information by facial landmark heatmaps into an end-to-end deep network. This pose information not only forms a mask image to guide the generator in learning process but also provides a flexible controllable condition during inference, which makes our CAPG-GAN differ from both previous GAN and 3DMM based methods. Four kinds of loss functions have been used in CAPG-GAN to synthesize photorealistic and identity preserving results. Extensive experimental results on face synthesis and face recognition demonstrate that our method not only presents compelling visual results but also facilitates to improve recognition accuracy of large pose face recognition.

6. Acknowledgement

This work is partially funded by National Natural Science Foundation of China (Grant No. 61622310, 61473289, 61427811) and Beijing Municipal Science and Technology Commission (Grant No.Z161100000216144). It is jointly supported by CRIPAC and Huawei Technologies Co., Ltd (Contract No.: YBN2017100047).

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. In *ICML*, 2017.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017.

[3] X. Chen, X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[4] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017.

[5] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *PR*, 66:144–152, 2017.

[6] C. Ferrari, G. Lisanti, S. Berretti, and A. D. Bimbo. Effective 3d based frontalization for unconstrained face recognition. In *ICPR*, 2016.

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

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[9] R. Gross, I.Matthews, J. Cohn, T.Kanade, and S.Baker. Multi-PIE. *Image Vis Comput*, 28(5):807–813, 2010.

[10] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007.

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

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[17] A. Jourabloo and X. Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. *IJCV*, 2017.

[18] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (SPAE) for face recognition across poses. In *CVPR*, 2014.

[19] M. Kan, S. Shan, and X. Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016.

[20] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. J. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In *CVPR*, 2015.

[21] I. Masi, S. Rawls, G. G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016.

[22] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.

[23] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[24] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *ICCV*, 2015.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[26] L. Song, M. Zhang, X. Wu, and R. He. Adversarial discriminative heterogeneous face recognition. In *AAAI*, 2018.

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[28] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[29] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv:1511.02683*, 2016.

[30] X. Wu, L. Song, R. He, and T. Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI*, 2018.

[31] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T. Liu. Dual supervised learning. In *ICML*, 2017.

[32] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T. K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015.

[33] J. Yang, S. E. Reed, M. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015.

[34] 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.

[35] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett.*, 23(10):1499–1503, 2016.

[37] J. Zhao, L. Xiong, K. Jayashree, J. Li, F. Zhao, Z. Wang, S. Pranata, S. Shen, S. Yan, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, 2017.

[38] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[39] 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.

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

[41] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014.