FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis

Yujun Shen[1], Ping Luo[1,3], Junjie Yan[2], Xiaogang Wang[1], Xiaoou Tang[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{sy116, pluo, xtang}@ie.cuhk.edu.hk, yanjunjie@sensetime.com, xgwang@ee.cuhk.edu.hk

Abstract

Face synthesis has achieved advanced development by using generative adversarial networks (GANs). Existing methods typically formulate GAN as a two-player game, where a discriminator distinguishes face images from the real and synthesized domains, while a generator reduces its discriminativeness by synthesizing a face of photo-realistic quality. Their competition converges when the discriminator is unable to differentiate these two domains.

Unlike two-player GANs, this work generates identity-preserving faces by proposing FaceID-GAN, which treats a classifier of face identity as the third player, competing with the generator by distinguishing the identities of the real and synthesized faces (see Fig.1). A stationary point is reached when the generator produces faces that have high quality as well as preserve identity. Instead of simply modeling the identity classifier as an additional discriminator, FaceID-GAN is formulated by satisfying information symmetry, which ensures that the real and synthesized images are projected into the same feature space. In other words, the identity classifier is used to extract identity features from both input (real) and output (synthesized) face images of the generator, substantially alleviating training difficulty of GAN. Extensive experiments show that FaceID-GAN is able to generate faces of arbitrary viewpoint while preserve identity, outperforming recent advanced approaches.

1. Introduction

Image generation has received much attention in recent years [7, 10]. Among them, synthesizing a face image of a different viewpoint but preserving its identity becomes an important task, owing to its wide applications in industry, such as video surveillance and face analysis.

Recently, this task has been significantly advanced by generative adversarial networks (GANs). For example, as shown in Fig.1 (a.1), previous methods [33, 28] are typically built upon the original GAN [9], which is formulated as a two-player game, including a discriminator and a generator, denoted as D and G respectively. In the conventional GAN, G employs a real image $\mathbf{x}^r$ as input and outputs a synthesized image $\mathbf{x}^s$, while D adopts these two images as inputs and outputs whether they are real or

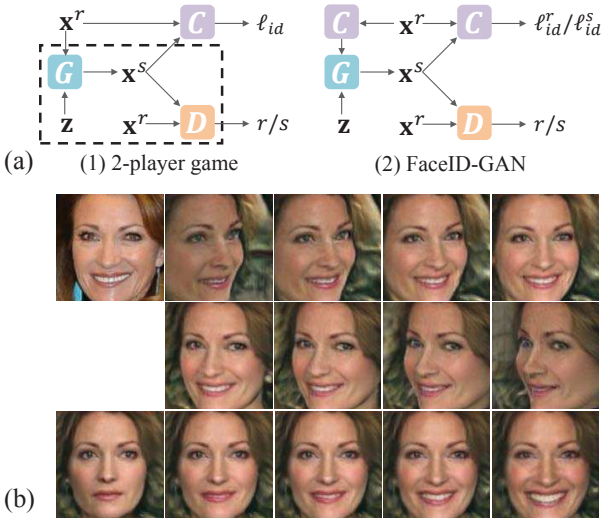

Figure 1. (a.1) shows that an original GAN (dashed box) is extended by an identity classifier C to predict identity label ℓ_{id}. It is formulated as a two-player game, where C does not compete with the generator G. G uses a real image $\mathbf{x}^r$ as input and outputs a synthesized image $\mathbf{x}^s$. $\mathbf{z}$ represents random noise. D is a discriminator to differentiate real and synthesized domains. (a.2) shows FaceID-GAN, which is a three-player GAN by treating C as the third player to distinguish identities of two domains, ℓ_{id}^r and ℓ_{id}^s. C collaborates together with D to compete with G, making G produce identity-preserving and high-quality images to confuse both C and D. FaceID-GAN is designed by using a criterion of information symmetry, where C is employed to learn idenity features for both domains. (b) visualizes some examples of FaceID-GAN, showing its capacity to generate faces of arbitrary viewpoint and expression, while preserving identity.

synthesized (fake). In training, D and G compete with each other, where the discriminator maximizes its classification accuracy, whereas the generator reduces accuracy of the discriminator by synthesizing images of high quality. Their competition converges when D is unable to distinguish the fake data from the real data, indicating that the qualities of images in these two domains are sufficiently close.

In order to produce identity-preserving face images, existing methods extend the original GAN by using an additional classifier, denoted as C, which employs both $\mathbf{x}^r$ and $\mathbf{x}^s$ as inputs, and predicts their identity labels, denoted as $\ell_{id} \in \mathbb{R}^{N\times 1}$. This label represents a 1-of-N vector of N subjects, where each entry indicates the probability of an image belonging to a certain subject. In other words, to preserve identity, G is expected to output a face (synthesized) with the same identity label with its cooresponding input (real) under the supervision of C, as shown in Fig.1 (a.1).

In the above, although C is able to learn identity features, it is unable to satisfy the requirement of preserving identity, *i.e*. to push real and synthesized domains as close to each other as possible. This is illustrated in Fig.2 (a). Given two real images that have different identities, with identity features $\mathbf{f}^r_{id1}$ and $\mathbf{f}^r_{id2}$, and a synthesized image, with identity feature $\mathbf{f}^s_{id1}$, which is expected to have the first identity. In previous approaches, when the distance between $\mathbf{f}^s_{id1}$ and $\mathbf{f}^r_{id1}$ is smaller than the distances between $\mathbf{f}^s_{id1}$ and all the remaining identities, *i.e*. $\mathbf{f}^s_{id1}$ locates next to the boundary but slightly biases towards $\mathbf{f}^r_{id1}$, C is sufficient to assign them the same label, but neglects how close they are in the feature space, impeding the identity-preserving capacity.

This work argues that building on the conventional two-player GAN as existing methods have done, is not sufficient to preserve face identity. To this end, we present FaceID-GAN, a novel deep generative adversarial network that is able to synthesize face images of arbitrary viewpoint, while well preserving identity as shown in Fig.1 (b). It has two appealing properties.

First, FaceID-GAN provides a novel perspective by extending the original two-player GAN to a GAN with three players. Unlike previous methods that treat C as a spectator, which does not compete with G, FaceID-GAN treats C as the third player, which not only learns identity features, but also differentiates two domains by assigning them different identity labels ℓ^r_{id} and ℓ^s_{id}, as shown in Fig.1 (a.2). Intuitively, in FaceID-GAN, C competes with G and cooperates with D. In particular, C and D distinguish two domains with respect to face identity and image quality respectively, whereas G tries to improve image generation to reduce their classification accuracies. Training is converged when C and D are unable to differentiate the two domains, implying that G is capable of producing face images that are photo-realistic as well as identity-preserving.

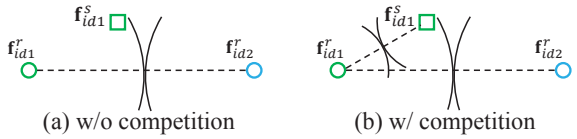

Figure 2. The merit of treating C as a competitor in FaceID-GAN.

Fig.2 (b) illustrates the merit of the above procedure, where the notations are similar as Fig.2 (a). In FaceID-GAN, C not only classifies between "id1" and "id2", but also between real "id1" and fake "id1", by using $2N$ labels. In this case, in order to confuse C, G has to synthesize an image, whose identity feature, $\mathbf{f}^s_{id1}$, is not only located inside the boundary of $\mathbf{f}^r_{id1}$, but also moved towards $\mathbf{f}^r_{id1}$ as much as possible, reducing the distance between them so as to decrease classification accuracy of C. After competition, G is able to substantially preserve face identity.

Second, this work designs FaceID-GAN by following information symmetry, which is a general principle to design the architectures of GANs. As shown in Fig.1 (a.2), C in FaceID-GAN extracts features from both $\mathbf{x}^r$ and $\mathbf{x}^s$, leading to symmetry of information, unlike (a.1) where identity feature of $\mathbf{x}^s$ is extracted by using C, but that of $\mathbf{x}^r$ is extracted by using G implicitly. Recall that the network has to move $\mathbf{f}^s_{id1}$ towards $\mathbf{f}^r_{id1}$ in attempt to preserve identity, as shown in Fig.2. If these features are extracted by using G and C separately, the distance between them is probably large, bringing training difficulty, because these two modules represent two different feature spaces. In contrast, since features of both domains are extracted by using C in FaceID-GAN, their distance could be close, even at the beginning when the network is trained from scratch, significantly reducing the training difficulty.

To summary, this work has three main **contributions**. (1) The conventional two-player GAN is extended to three players in FaceID-GAN, where the identity classifier collaborates together with the discriminator to compete with the generator, producing face images that have high quality and well preserved identity. (2) To design FaceID-GAN, we present information symmetry to alleviate training difficulty. It can be treated as a general principle to design the architectures of GANs. (3) Besides high visual quality, FaceID-GAN is able to generate face images with high diversity in viewpoints and expressions, surpassing the recent advanced methods, which are carefully devised to deal with pose variations.

2. Relations to Previous Work

This section summarizes previous works that synthesize face images by using deep generative models, and compares them to FaceID-GAN. In the literature, there are many methods of image generation that do not employ GANs. We would also like to acknowledge their contributions [12, 39, 40, 32].

In general, existing deep models can be categorized into three groups, based on their learned input-output image mappings, including one-to-one, many-to-one, and many-to-many as shown in Fig.3, where different networks have different components. Besides those mentioned before, E denotes an encoder that projects a real image into a hidden feature space, and P is a facial shape feature extractor. In this part, we just take viewpoints (poses) as an example.

One-to-one. Some works learn one-to-one mapping as shown in Fig.3 (a), where a face of one style is transformed to the other, such as from image to sketch [34], from low to high resolution [8], and from visible to infrared spectrum [20]. In the early stage, these tasks were often solved by using encoder-decoder structures, where E encodes $\mathbf{x}^r$ to hidden feature $\mathbf{h}$, G transforms this feature to $\mathbf{x}^s$, and C predicts identity. In this setting, as shown by the red arrows in (a), G is trained by minimizing the per-pixel difference between $\mathbf{x}^s$ and its ground truth image ℓ_I^s.

Many-to-one. With GANs [9, 1, 36], the network in (a) is extended to learn many-to-one mapping as shown in (b), such as face frontalization [6, 15, 33], which transforms multiview to frontal view. With the conventional GANs, G and D are two competitors, while C is a spectator that learns facial identity. However, in this setting, since the input data distribution has larger variations (multiview) than that of the output (single view), the pose label of the real image, ℓ_p^r, is employed as conditional input to reduce training difficulty.

The above methods require the ground truth image ℓ_I^s as supervision, and the label ℓ_p^r as input, impeding their applications in a more challenging setting as shown in (c). Given an image of arbitrary pose as input, the network in (c) produces faces of different poses, while preserving identity [16, 28, 29]. This problem is extremely challenging, because both the input and output data distributions have multiple modes.

Many-to-many. There are three major improvements when comparing (c) to (b). First, a module of pose P is used as a constraint, ensuring $\mathbf{x}^s$ had the desired pose ℓ_p^s. Second, G has three inputs rather than two, where $\mathbf{h}$ is the same but the other two are different, including a vector of random noise, $\mathbf{z}$, and the desired pose, ℓ_p^s. In fact, the network cannot be trained without them. For example, if $(\mathbf{x}^r, \ell_p^r)$ are fed into G just like what (b) does, the network fails to produce $\mathbf{x}^s$ of different poses, because transforming the same input to multiple outputs has large ambiguity. Instead, $(\mathbf{x}^r, \ell_p^s, \mathbf{z})$ are used to reduce ambiguity and improve diversity of the generated images. Third, the ground truth image, ℓ_I^s, is removed and the per-pixel loss between $\mathbf{x}^s$ and ℓ_I^s is also removed, enabling training with unpaired data. In other words, the network learns to generate $\mathbf{x}^s$ of different poses, no matter whether the corresponding ground truth image exists or not.

Although these methods eliminate the paired data assumption, they still rely on label of pose, limiting their generalization capacity. For example, since the label is defined as a discrete 1-of-K vector, it is difficult to generalize to a full spectrum of viewpoints, which is smooth and continuous. Furthermore, as aforementioned, the methods in (a-c) break the symmetry of information. For example, in (c), $\mathbf{x}^r$ is represented by $\mathbf{h}$ extracted by E, while $\mathbf{x}^s$ is represented by $\mathbf{f}_{id}^s$ extracted by C. Obviously, their identity information are represented by features of different spaces. In other words, before learning to produce $\mathbf{x}^s$ with preserved identity, G is required to learn a transition between feature spaces, *e.g.* from $\mathbf{h}$ to $\mathbf{f}_{id}^s$, bringing non-negligible training difficulty.

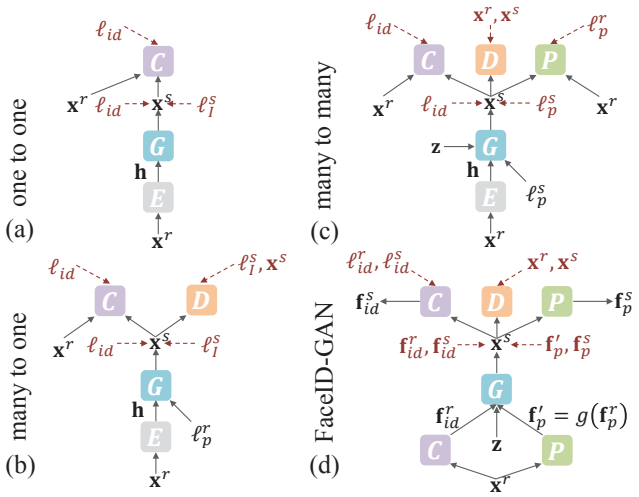

Figure 3. FaceID-GAN is compared to existing works, including learning (a) one-to-one, (b) many-to-one, and (c) many-to-many mappings for face generation. In all these four figures, the arrows in dark represent forward computations, whilst the dashed arrows in red represent backward supervisions. Better viewed in color and zoom in 150%.

FaceID-GAN. In (d), FaceID-GAN addresses these weaknesses in two folds. First, all above methods are based on two-player GANs, which have flat performances in preserving identity. In contrast, FaceID-GAN introduces a third player, making G to compete with C and D simultaneously, and hence synthesize faces with both preserved identity and high quality. Second, FaceID-GAN follows information symmetry criterion by replacing E with P and C, where parameters of the two modules of P (or C) are shared. In this case, $\mathbf{x}^r$ and $\mathbf{x}^s$ are represented by using the same feature space, alleviating the training difficulty. Moreover, G is directly supervised by features from two domains instead of by certain labels, leading to a better generation.

In this work, we use the well-known 3D Morph Model (3DMM) [3] to help represent the facial shape feature $\mathbf{f}_p$, including not only pose, but expression and general shape information as well.

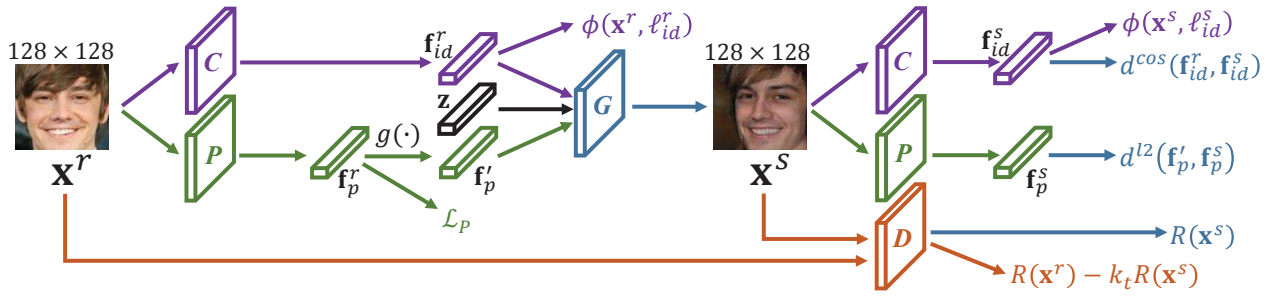

Figure 4. The overall framework of FaceID-GAN. G is the generator that collects all information and synthesizes faces under certain constraints. P is the facial shape estimator to provide shape information. C provides identity information and also competes with G with respect to facial identity. D is the discriminator that competes with G from quality aspect. Better viewed in color.

3. Proposed Method

Fig.4 illustrates the overall framework of FaceID-GAN. Given an input face $\mathbf{x}^r$ of 128×128, P estimates its facial shape feature, $\mathbf{f}_p^r \in \mathbb{R}^{229}$, based on the 3D Morphable Model [3], and then turns $\mathbf{f}_p^r$ into the desired shape feature, $\mathbf{f}_p'$, by using a transformation denoted as $g(\cdot)$. We have $\mathbf{f}_p' = g(\mathbf{f}_p^r)$, which represents the desired pose and expression. C is a face recognition module to extract identity feature, $\mathbf{f}_{id}^r \in \mathbb{R}^{256}$. G employs $\mathbf{f}_p'$, $\mathbf{f}_{id}^r$, and a random noise $\mathbf{z} \in \mathbb{R}^{128}$ as inputs, and synthesizes a face image $\mathbf{x}^s$ of size 128×128, denoted as $\mathbf{x}^s = G(\mathbf{f}_p', \mathbf{f}_{id}^r, \mathbf{z})$.

Overview. As discussed before, FaceID-GAN has three players, including D, C, and G, where the first two cooperate to discriminate real and synthesized domains, while the last one reduces their discriminateness. In this work, G is also supervised by P, so as to control viewpoint and expression. The loss functions for the above components are defined as

$$\min_{\Theta_D} \mathcal{L}_D = R(\mathbf{x}^r) - k_t R(\mathbf{x}^s), \tag{1}$$

$$\min_{\Theta_C} \mathcal{L}_C = \phi(\mathbf{x}^r, \ell_{id}^r) + \lambda \phi(\mathbf{x}^s, \ell_{id}^s), \tag{2}$$

$$\min_{\Theta_G} \mathcal{L}_G = \lambda_1 R(\mathbf{x}^s) + \lambda_2 d^{cos}(\mathbf{f}_{id}^r, \mathbf{f}_{id}^s) + \lambda_3 d^{l2}(\mathbf{f}_p', \mathbf{f}_p^s), \tag{3}$$

where $R(\cdot)$, $\phi(\cdot,\cdot)$, $d^{cos}(\cdot,\cdot)$, and $d^{l2}(\cdot,\cdot)$ denote different energy functions. λ, λ_1, λ_2, and λ_3 are different weight parameters between these functions. k_t is a regularization coefficient that balances between $R(\mathbf{x}^r)$ and $R(\mathbf{x}^s)$ at the t-th update step. Intuitively, $\mathcal{L}_D$ minimizes the energy of $R(\mathbf{x}^r)$ but maximizes that of $R(\mathbf{x}^s)$ to distinguish two domains according to their image quality. $\mathcal{L}_C$ contains identity classifier ϕ, which classifies faces of two domains by using different identity labels as introduced before, to differentiate two domains according to their identities. In $\mathcal{L}_G$, G tries to compete with D by producing high-quality face to minimize $R(\mathbf{x}^s)$. G also is trained to minimize the cosine distance between identity features of $\mathbf{x}^r$ and $\mathbf{x}^s$, denoted as $d^{cos}(\mathbf{f}_{id}^r, \mathbf{f}_{id}^s)$, and to minimize l_2 euclidean distance between the synthesized shape feature $\mathbf{f}_p^s$ and the desired shape feature $\mathbf{f}_p'$, denoted as $d^{l2}(\mathbf{f}_p', \mathbf{f}_p^s)$, in order to preserve identity and change pose and expression. We discuss details of these components in the following.

3.1. Discriminator D

The conventional way [9] to to discriminate the real and synthesized domains is by using a binary classifier. However, it is infeasible for image generation because of its sparse supervision. To incorporate per-pixel supervisions [1, 35], this work employs an auto-encoder as discriminator D, which is introduced in [2]. In other words, D works on reconstructing an input image by minimizing the pixel-wise distance between input and output. We have $R(\mathbf{x}) = ||\mathbf{x} - D(\mathbf{x})||_1$ and $\mathcal{L}_D = R(\mathbf{x}^r) - k_t R(\mathbf{x}^s)$, which differentiates two domains by minimizing the reconstruction error of real image, but maximizing that of synthesized image.

Following [2], to keep balance between $R(\mathbf{x}^r)$ and $R(\mathbf{x}^s)$, we introduce a regularization term k_t, which is dynamically updated in the training process,

$$k_{t+1} = k_t + \lambda_k(\gamma R(\mathbf{x}^r) - R(\mathbf{x}^s)), \tag{4}$$

where λ_k is the learning rate and γ represents a diversity ratio of $\mathbf{x}^s$. We set $\lambda_k = 0.001$ and $\gamma = 0.4$ in this work.

3.2. Classifier C

To retain identity, C learns identity features of the real and synthesized images, $\mathbf{x}^r$ and $\mathbf{x}^s$, whose features are denoted as $\mathbf{f}_{id}^r$ and $\mathbf{f}_{id}^s$. C discriminates two domains by classifying them using different labels. We formulate C as a 1-of-$2N$ classification problem, with the purpose of classifying $\mathbf{x}^r$ to first N labels and $\mathbf{x}^s$ to the last N labels, by using the cross-entropy loss

$$\begin{aligned} \phi(\mathbf{x}^r, \ell_{id}^r) &= \sum_j -\{\ell_{id}^r\}_j \log(\{C(\mathbf{x}^r)\}_j), \\ \phi(\mathbf{x}^s, \ell_{id}^s) &= \sum_j -\{\ell_{id}^s\}_j \log(\{C(\mathbf{x}^s)\}_j), \end{aligned} \tag{5}$$

where $j \in [1, 2N]$ is the j-th index of identity classes. However, treating these $2N$ classes equally is unreasonable, because in the identity feature space, a synthesized face should be closer to its corresponding real input face when comparing with the other identities. Therefore, we introduce a loss weight λ as shown in Eqn.(2), to balance the contribution of the synthesized faces, making C learn more accurate identity representation.

3.3. Shape Estimator P

We incorporate the 3D Morphable Model (3DMM) [3] to project facial images into a shape feature space, representing pose and expression. The 3DMM of faces is formulated as

$$\begin{aligned} &\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}, \\ &\mathbf{V}(\mathbf{p}) = f * \mathbf{R}(\theta_x, \theta_y, \theta_z) * \mathbf{S} + [t_x, t_y, t_z]^T, \\ &\mathbf{p} = [{\boldsymbol{\alpha}_{id}}^T, {\boldsymbol{\alpha}_{exp}}^T, \theta_x, \theta_y, \theta_z, t_x, t_y, t_z]^T, \\ &\mathbf{f}_p = [{\boldsymbol{\alpha}_{id}}^T, {\boldsymbol{\alpha}_{exp}}^T, \theta_y]^T, \end{aligned} \tag{6}$$

where $\overline{\mathbf{S}}$ is a mean shape of a 3D face, $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ are the PCA bases for shape [23] and expression [4] respectively. Therefore, $\mathbf{S}$ is the 3D shape of certain shape and expression in the 3DMM coordinate system, which can be uniquely defined by the coefficients $\boldsymbol{\alpha}_{id} \in \mathbb{R}^{199}$ and $\boldsymbol{\alpha}_{exp} \in \mathbb{R}^{29}$. Here $\mathbf{V}(\cdot)$ is a 3D shape in the image coordinate system, which is obtained by transforming $\mathbf{S}$ using scaling coefficient f, rotation coefficients $[\theta_x, \theta_y, \theta_z]^T$, and translation coefficients $[t_x, t_y, t_z]^T$. $\mathbf{p}$ denotes the complete set of parameters in 3DMM. Among these parameters, $\boldsymbol{\alpha}_{id}$ provides the general shape information, which differs between identities, while $\boldsymbol{\alpha}_{exp}$ and θ_y control expression and pose respectively.

To achieve end-to-end training, we incorporate a network P to learn the shape feature. Before training, we follow [38] to prepare the 3DMM coefficients for all the images $\mathbf{x}^r$, denoted as $\overline{\mathbf{f}}_p^r$. Similar to [37], P is trained to minimize the weighted distance function

$$\min_{\Theta_P} \mathcal{L}_P = (P(\mathbf{x}^r) - \overline{\mathbf{f}}_p^r)^T \mathbf{W} (P(\mathbf{x}^r) - \overline{\mathbf{f}}_p^r), \tag{7}$$

where $\mathbf{W}$ is an importance matrix whose diagonal elements are the weights.

Different from C that preserves identity of two domains, this work requires pose and expression to be varied but not preserved. Therefore, given an input real image $\mathbf{x}^r$, we use P to extract its shape feature $\mathbf{f}_p^r = [{\boldsymbol{\alpha}_{id}^r}^T, {\boldsymbol{\alpha}_{exp}^r}^T, \theta_y^r]^T$ and transform this feature by using

$$\mathbf{f}_p' = g(\mathbf{f}_p^r, \boldsymbol{\alpha}_{exp}', \theta_y') = [{\boldsymbol{\alpha}_{id}^r}^T, {\boldsymbol{\alpha}_{exp}'}^T, \theta_y']^T, \tag{8}$$

where $\boldsymbol{\alpha}_{id}^r$ represents the original shape information of $\mathbf{x}^r$, and $\boldsymbol{\alpha}_{exp}'$ and θ_y' represent the desired pose and expression.

In other words, we disentangle $\mathbf{f}_p^r$ into facial shape, pose, and expression by using the network P, and use function $g(\cdot)$ to introduce randomness to shape feature. Meanwhile, G is trained to generate faces under the supervision of P by keeping $\mathbf{f}_p'$ remained, no matter what $\boldsymbol{\alpha}_{exp}'$ and θ_y' are. In this way, G is able to synthesis faces of arbitrary pose and expression.

3.4. Generator G

As shown in Eqn.(3), besides minimizing $R(\mathbf{x}^s)$ to compete with D, G is also trained by minimizing two distances

$$d^{cos}(\mathbf{f}_{id}^r, \mathbf{f}_{id}^s) = 1 - \frac{{\mathbf{f}_{id}^r}^T \mathbf{f}_{id}^s}{||\mathbf{f}_{id}^r||_2 ||\mathbf{f}_{id}^s||_2}, \tag{9}$$

$$d^{l2}(\mathbf{f}_p', \mathbf{f}_p^s) = ||\mathbf{f}_p^s - \mathbf{f}_p'||_2^2, \tag{10}$$

where identity is preserved by minimizing the cosine distance between identity features of the real and synthesized images, whereas pose and expression are changed by minimizing the euclidian distance between shape feature of the synthesized face and the desired shape.

3.5. Implementation Details

Before training, all faces used in this work are aligned by using [26] to image size 128×128. C employs ResNet-50 [13] by changing the active function from "BN+ReLU" to "SELU" [18]. P employs ResNet-18, on top of which we add three more fully-connected layers following [24]. G and D use BEGAN structure [2].

At the training stage, $\mathbf{z}$ is sampled from a uniform distribution with range $[-1, 1]$. λ_1, λ_2, λ_3 are set to balance the initial losses of generator G, corresponding to D, C and P respectively. λ is 1 at the beginning and gradually decreases as training proceeds. The batch size is 96, equally distributed to 8 GPUs. We use Adam optimizer [17] for all four components, and the parameters are updated for $200k$ steps. The initial learning rate (lr) of G and D is 0.0008, and drops 0.0002 for every $50k$ steps. The initial lr of C is 0.0008, and drops to 0.0005 at $150k$-th step. Except P is well pre-trained, all other three modules in this work are trained from scratch.

4. Experiments

FaceID-GAN aims at synthesizing high-quality identity-preserving faces, but with high diversity in poses and expressions. Extensive experiments are conducted to compare FaceID-GAN with existing face generation methods, including face synthesis and face verification.

Datasets. FaceID-GAN is trained on CASIA-WebFace [31] and evaluated on multiple different datasets, including LFW [14], IJB-A [19], CelebA [21], and CFP [25]. We briefly introduce these datasets in the following.

1) CASIA-WebFace. It consists of $494,414$ images of $10,575$ subjects. We employ it for training. **2) LFW.** It consists of $13,233$ images of $5,749$ subjects. With LFW, existing works [38, 33] evaluate their performances on the tasks of face frontalization and verification. FaceID-GAN is also applicable to these tasks by generating faces with frontal viewpoint and neutral expression, though it is not specially designed for this purpose. We evaluate FaceID-GAN on these tasks following existing protocols and compare it to previous works. **3) IJB-A.** It consists of $25,808$ images of 500 subjects. Following previous works, we evaluate the identity-preserving capacity of FaceID-GAN on this dataset. **4) CelebA.** This is a large-scale dataset that contains $202,599$ images of $10,177$ subjects, where the face images have large diversity, making it an appropriate test set for face image synthesis. **5) CFP.** It consists of 500 subjects, each of which has 10 images in frontal view and 4 images in profile view. Following prior work [29], we evaluate the effectiveness of FaceID-GAN to generate faces under different viewpoints.

The above evaluation sets (2-5) consist of nearly 250K images from a wide spectrum of viewpoints and subjects. The overlapping ratio between the subjects in training and evaluation sets is smaller than 0.1%. The experimental setting used in this work is challenging, substantially characterizing the superiorities over existing algorithms.

4.1. Face Synthesis

This section evaluates FaceID-GAN from three aspects, including image quality, control of pose and expression, and ability to preserve identity.

Frontalization and Identity. Generating faces of canonical viewpoint is an important task, because it reduces facial variations that hinder face verification. Unlike previous approaches that are specially designed to address this problem, such as HPEN[38] and FF-GAN [33], FaceID-GAN treats it as a subtask by synthesizing faces with the desired pose of $0°$. Fig.5 visualizes results of four images selected from LFW by following [33]. On the left corner of each synthesized image is a score, which indicates the similarity of identity between the input and output images. To compute these scores, we employ a face recognition model [27] trained on MS-Celeb-1M [11], which is totally independent of this work, making the results convincing. This model is also applied to the remaining experiments.

From Fig.5, we see that faces generated by FaceID-GAN outperform others in following aspects. First, FaceID-GAN produces faces of exactly frontal viewpoint, which benefits from the shape controlling module P, while previous methods produced distortions. Second, FaceID-GAN generates a new face, instead of just learning an interpolation. This is because we filter out extraneous information by passing identity feature into generator instead of the full image.

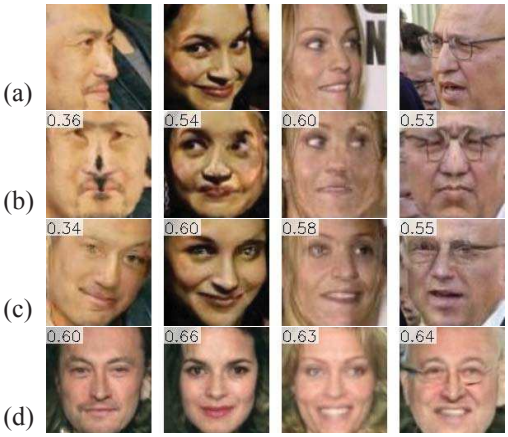

Figure 5. Face frontalization results on LFW. (a) Input. (b) HPEN [38]. (c) FF-GAN [33]. (d) FaceID-GAN (ours). On the top-left corner of each frontalized face, a score indicates the identity similarity between the input and the generated face.

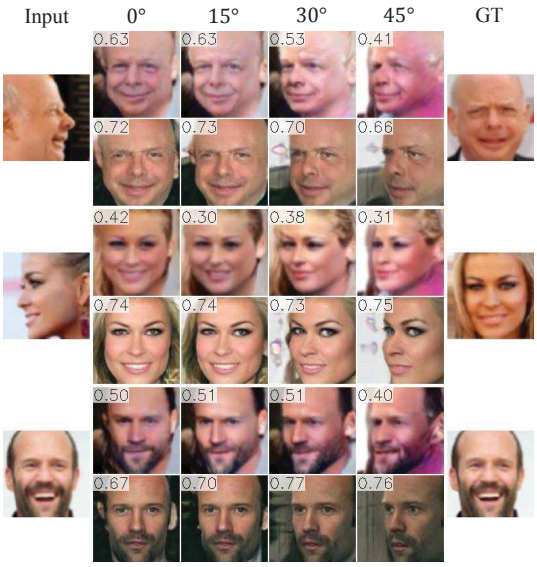

Figure 6. Face rotation results on CFP. Odd rows are the rotation results from DR-GAN [29], and even rows are results from FaceID-GAN (ours).

Third, FaceID-GAN makes a better synthesis from the view of both image quality and identity maintenance, owing to the three-player competition.

Rotation and Identity. In this part, we evaluate the effectiveness of FaceID-GAN when generating faces under different viewpoints, while maintaining face identity. We compare our method with DR-GAN [29], which is a recent advanced method to solve this task. Both methods are trained on CASIA-WebFace and then directly evaluated on CFP without fine-tuning the models on the CFP dataset. We select a set of testing images that are the same as [29]. Fig.6 shows the comparison results by rotating face from $0°$ to $45°$. In this case, the identity similarity is computed by using ground truth image as reference.

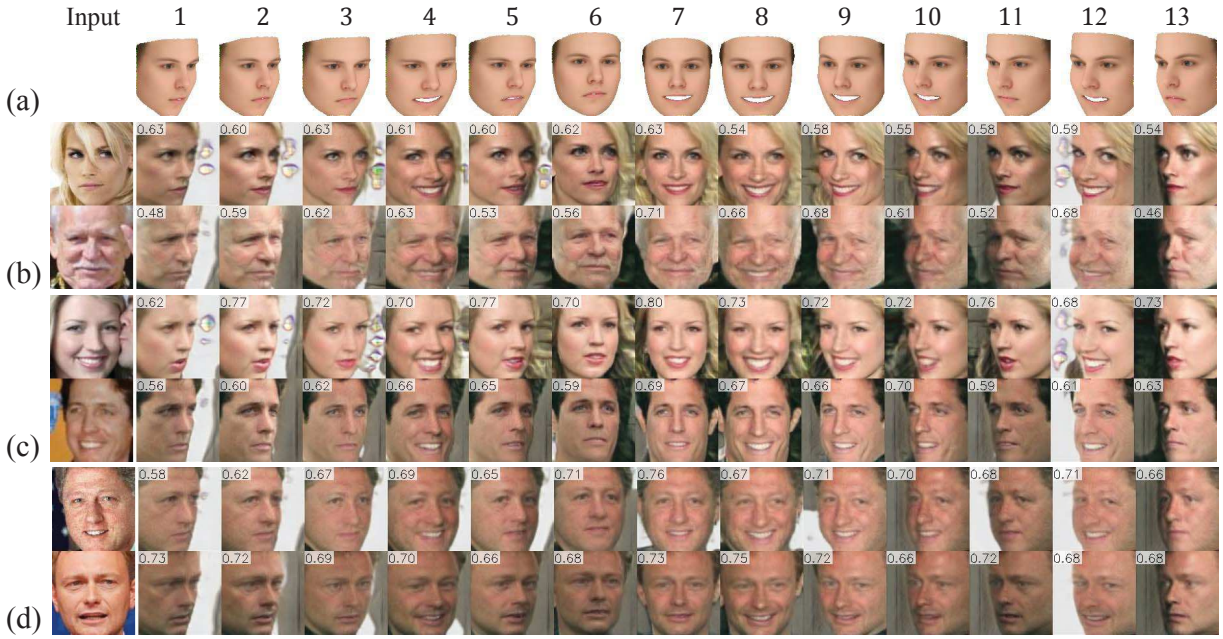

Figure 7. Face synthesis results of different datasets, including (b) CelebA, (c) LFW, and (d) IJB-A. (a) shows the 3D templates with the desired poses and expressions. In (b-d), the first column is the input image, whilst the remaining columns show the synthesized results.

As shown in Fig.6, several evidences suggest that our method has a strong ability for face synthesis. First, FaceID-GAN produces faces with higher resolution than DR-GAN, which is trained on images of 96×96, but FaceID-GAN is able to train with 128×128 face images. Second, FaceID-GAN preserves identity much better than DR-GAN, especially when large pose is presented. Third, in DR-GAN, the image quality drops rapidly when rotating the faces, while FaceID-GAN demonstrates stableness and robustness. Finally, besides the angles reported in Fig.6, FaceID-GAN can actually rotate face with arbitrary angle as shown in Fig.7.

Pose, Expression and Identity. We further evaluate the generalization capacity of FaceID-GAN by controlling pose and expression, and maintaining identity simultaneously. Fig.7 visualizes results on multiple datasets, including CelebA, LFW, and IJB-A. Note that we do not fine-tune the model to adapt these datasets.

In Fig.7 (a), we illustrate the 3D templates of the desired poses and expressions. From (b) to (d), we see that the synthesized images exactly match the templates in (a), while preserving the face identity. We also observe that the face characteristics can be preserved for different identities. For example, under the 8-th face template, smiling is presented differently in different identities, showing that FaceID-GAN can learn to change expression while preserving identity's characteristics. However, synthesized faces don't vary much in expressions. This is attributed to the fact that most faces in training set are with neutral or smile expression. Just like other GAN-based framework, FaceID-GAN generates images by learning the underlying distribution of input data. A training set with larger diversity in expressions will likely alleviate this issue.

In summary, FaceID-GAN can synthesize high-quality facial images by maintaining identity and controlling pose and expression.

4.2. Face Verification

Here we evaluate FaceID-GAN on LFW and IJB-A for face verification, to further verify its identity preserving ability. The identity features of both real and synthesized images are extracted by module C in FaceID-GAN and the cosine distance is used as the metric for face verification.

Evaluation on LFW. By following existing works [12, 38, 33], we frontalize all the images in LFW and then evaluate face verification performance on the frontalized images, to verify whether FaceID-GAN retains the identity information properly. Tab.1 compares our results to the state-of-the-art methods. The improvement demonstrates that FaceID-GAN can synthesis both realistic and identity-preserving face images.

Evaluation on IJB-A. We further evaluate the verification performance on IJB-A dataset. IJB-A defines a different protocol by matching templates, where each template contains a variant amount of images. We define a confidence value of each image to be the reciprocal of its corresponding reconstruction error estimated by D. This value can describe image quality to some extent. We fuse the features of images in a template with their confidence values as weights, and use the fused result to represent such template.

Tab.2 shows the verification accuracy on IJB-A. Comparing to the state-of-art methods, FaceID-GAN achieves superior results both at FAR 0.01 and at FAR 0.001, which suggests that by competing with G, C and D in FaceID-GAN also perform well in their respective tasks, *i.e*. identity reprensentation learning and image quality evaluation.

4.3. Ablation Study

FaceID-GAN consists of four components, C, P, D and G, where G is necessary for a generative model. To evaluate the contributions of other three parts, we train three models by removing these components respectively, while keeping the training process and all hyper-parameters the same. Among them, the model without D diverges. Meanwhile, FaceID-GAN advances existing methods by proposing 3-player competition and information symmetry, so we train two extra models to evaluate these two improvements. Among them, the convergence of the model without following information symmetry is slow and instable.

Fig.8 shows the visual results generated by remaining three models as aforementioned, as well as the full model. Our proposed FaceID-GAN outperforms the others in the following aspects: visual effect (image quality), identity preservation, and the capability to control pose and expression. For example, the images in (b) have exactly the same facial shape with the inputs, demonstrating the ability of 3DMM to represent pose and expression. However, only providing shape information to the generator is not sufficient to preserve identity. This problem is greatly alleviated with the help of a face recognition engine. Images in (c) show better results by retaining more identity information. But it still fails when dealing with large poses, and the pose of synthesized faces becomes uncontrollable. Images in (d) are generated by incorporating the elements of both (b) and (c). By comparing (c) and (d), we conclude that besides controlling pose, 3DMM also plays an important role by providing general facial shape information, especially for inputs with extreme poses. The comparison of (d) and (e) shows that by introducing the third player, FaceID-GAN achieves better synthesis from the aspects of both image quality and identity preservation. To better illustrate the improvement from (d) to (e), we pick 10,000 face pairs, which are generated by (d) and (e) respectively, and ask annotators to vote which one is better in terms of visual quality and similarity of identity. As a result, (e) beats (d) with 56% votes for higher quality and 72% for better preserving identity.

5. Conclusion

In this work, we propose an end-to-end deep framework, FaceID-GAN, with the ability to synthesize photo-realistic face images of arbitrary viewpoint and expression, while preserving face identity. FaceID-GAN is formulated as a three-player GAN by introducing an identity classifier C as an additional competitior to the conventional GAN. C cooperates together with the discriminator D to compete with the generator G from two different aspects, facial identity and image quality respectively. An information symmetry criterion is also presented to design the architecture of FaceID-GAN, improving the performance and stability by alleviating training difficulty. We believe this work is promising as a general method for effectively solving other conditional generative problems.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (No. 61503366), the Big Data Collaboration Research grant from SenseTime Group (CUHK Agreement No. TS1610626), and the General Research Fund (GRF) of Hong Kong (No. 14236516).

Table 1. Performance comparison on LFW

Method	Verification Accuracy
3D [12]	93.62 ± 1.17
HPEN [38]	96.25 ± 0.76
FF-GAN [33]	96.42 ± 0.89
FaceID-GAN (ours)	$\mathbf{97.01} \pm 0.83$

Table 2. Performance comparison on IJB-A

	Verification Accuracy	
Method	@FIR=0.01	@FIR=0.001
Wang *et al*. [30]	72.9 ± 3.5	51.0 ± 6.1
PAM [22]	73.3 ± 1.8	55.2 ± 3.2
DCNN [5]	78.7 ± 4.3	–
DR-GAN [28]	77.4 ± 2.7	53.9 ± 4.3
FF-GAN [33]	85.2 ± 1.0	66.3 ± 3.3
FaceID-GAN (ours)	$\mathbf{87.6} \pm 1.1$	$\mathbf{69.2} \pm 2.7$

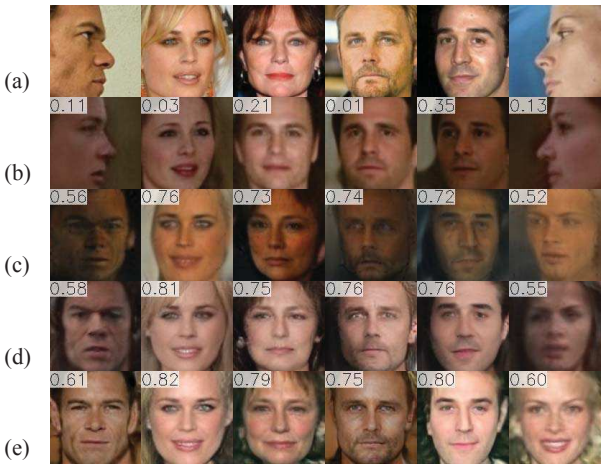

Figure 8. Visual results from abliation study. (a) Input. (b) FaceID-GAN *w/o* C. (c) FaceID-GAN *w/o* P. (d) FaceID-GAN *w/o* the competition between G and C. (e) FaceID-GAN (ours).

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3, 4

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 4, 5

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 3, 4, 5

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 2014. 5

[5] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *WACV*, 2016. 8

[6] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017. 3

[7] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 1995. 1

[8] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 3

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 3, 4

[10] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 1

[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. 6

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 2, 7, 8

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[14] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 5

[15] 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. *arXiv preprint arXiv:1704.04086*, 2017. 3

[16] Y. Huang and S. M. Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *CVPRW*, 2017. 3

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *arXiv preprint arXiv:1706.02515*, 2017. 5

[19] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 5

[20] V. A. Knyaz, O. Vygolov, V. V. Kniaz, Y. Vizilter, V. Gorbatsevich, T. Luhmann, N. Conen, W. Forstner, K. Khoshelham, S. Mahendran, et al. Deep learning of convolutional auto-encoder for image matching and 3d object reconstruction in the infrared range. In *CVPR*, 2017. 3

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[22] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016. 8

[23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 5

[24] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017. 5

[25] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 5

[26] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 5

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 6

[28] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 1, 3, 8

[29] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *arXiv preprint arXiv:1705.11136*, 2017. 3, 6

[30] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. In *ICB*, 2015. 8

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[32] 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. 2

[33] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 1, 3, 6, 7, 8

[34] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *ICMR*, 2015. 3

[35] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 4

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 3

[37] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 5

[38] 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. 5, 6, 7, 8

[39] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013. 2

[40] 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. 2