Disentangled Representation Learning GAN for Pose-Invariant Face Recognition

Luan Tran, Xi Yin, Xiaoming Liu
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{tranluan, yinxi1, liuxm}@msu.edu

Abstract

The large pose discrepancy between two face images is one of the key challenges in face recognition. Conventional approaches for pose-invariant face recognition either perform face frontalization on, or learn a pose-invariant representation from, a non-frontal face image. We argue that it is more desirable to perform both tasks jointly to allow them to leverage each other. To this end, this paper proposes Disentangled Representation learning-Generative Adversarial Network (DR-GAN) with three distinct novelties. First, the encoder-decoder structure of the generator allows DR-GAN to learn a generative and discriminative representation, in addition to image synthesis. Second, this representation is explicitly disentangled from other face variations such as pose, through the pose code provided to the decoder and pose estimation in the discriminator. Third, DR-GAN can take one or multiple images as the input, and generate one unified representation along with an arbitrary number of synthetic images. Quantitative and qualitative evaluation on both controlled and in-the-wild databases demonstrate the superiority of DR-GAN over the state of the art.

1. Introduction

Face recognition is one of the most widely studied topics in computer vision. Recently, great progress is achieved with Deep Learning-based methods [28, 33]. For example, suppressing-human performance is reported by Schroff et al. [33] on LFW database, which consists of mostly near-frontal faces. However, Pose-Invariant Face Recognition (PIFR) is far from solved [2, 8, 21, 22]. A recent study [34] shows that the performance of most algorithms degrades over 10% from frontal-frontal to frontal-profile face verification, while human performance only drops slightly. This indicates that the pose variation remains to be a significant challenge in face recognition and warrants future study.

Existing PIFR methods can be grouped into two categories. First, some work employ face frontalization [11, 14, 31, 40, 43, 45] to synthesize a frontal face, where traditional face recognition methods are applicable. The ability to generate a realistic frontal face is also beneficial for law enforcement practitioners to identify suspects. Second, other work focus on learning discriminative features directly from the non-frontal faces through either one joint model [28, 33] or multiple pose-specific models [7, 25]. In contrast, we propose a novel framework to take the best of both worlds — *simultaneously learn pose-invariant identity representation* and *synthesize faces with arbitrary poses.*

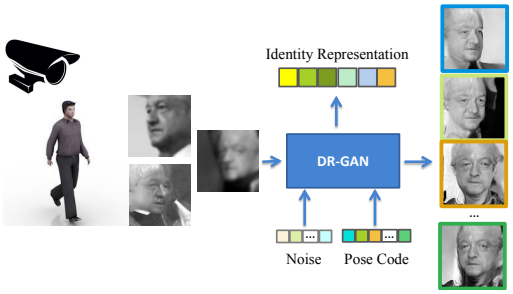


Figure 1: With one or multiple face images as the input, DR-GAN can produce an identity representation that is both discriminative and generative, i.e., the representation demonstrates superior PIFR performance, and can synthesize identity-preserving faces at target poses specified by the pose code.

As shown in Fig. 1, we propose Disentangled Representation learning-Generative Adversarial Network (DR-GAN) for PIFR. GAN [9] can generate samples similar to a data distribution through a two-player game between a generator G and a discriminator D. Despite many promising developments [5, 23, 26], image synthesis remains the main objective of GAN. Motivated by this objective, and the desire to learn an identity representation for PIFR, we construct G with an encoder-decoder structure (Fig. 2 (d)). The input to the encoder G_{enc} is a face image of any pose, the output of the decoder G_{dec} is a synthetic face at a target pose, and the learnt representation bridges G_{enc} and G_{dec}. While G serves as a face rotator, D is trained to not only distinguish real vs. synthetic images, but also predict the identity and pose of a face. With the additional classifications, D strives for the rotated face to have the same identity as the input real face, which has two effects on G: 1) The rotated

face looks more like the input subject in terms of identity. 2) The learnt representation is more *inclusive* or *generative* for synthesizing an identity-preserving face.

In conventional GANs, G takes a random noise vector to synthesize an image. In contrast, our G takes a face image, a pose code $\mathbf{c}$, and a random noise vector $\mathbf{z}$ as the inputs, with the goal of generating a face of the same identity with the target pose that can fool D. Specifically, G_{enc} learns a mapping from the input image to a feature representation. The representation is then concatenated with the pose code and the noise vector to feed to G_{dec} for face rotation. The noise models facial appearance variations other than identity or pose. DR-GAN can learn a disentangled identity representation that is *exclusive* or *invariant* to pose and other variations, which is ideal for PIFR when achievable.

Most existing face recognition algorithms only take one image for testing. In practice, there are scenarios when a set of test images is available [16]. In this case, prior work fuse the results either in the feature level [3] or the distance-metric level [36]. Different from prior work, our fusion is conducted within a unified framework. Specifically, G_{enc} is trained to take multiple images as the inputs and produce the identity representation and a coefficient for each image. Using the dynamically learned coefficients, the representations of all images are combined as one representation. During testing, G_{enc} takes any number of images and generates a single identity representation. G_{dec} synthesizes a face at the specified pose using this representation and the pose code.

This paper makes the following contributions. 1) We propose DR-GAN via an encoder-decoder structured generator that can frontalize or rotate a face with an arbitrary pose, even the extreme profile. 2) Our representation learning is explicitly disentangled from the pose variation through the pose code in G and the pose estimation in D. 3) We propose a novel scheme to adaptively fuse multiple faces to a single representation based on the learnt coefficients. It fulfills the recognition need of surveillance face snapshots and enables the matching of two face sets. 4) We achieve state-of-the-art face recognition performance on Multi-PIE [10], CFP [34], and IJB-A [16] databases.

2. Prior Work

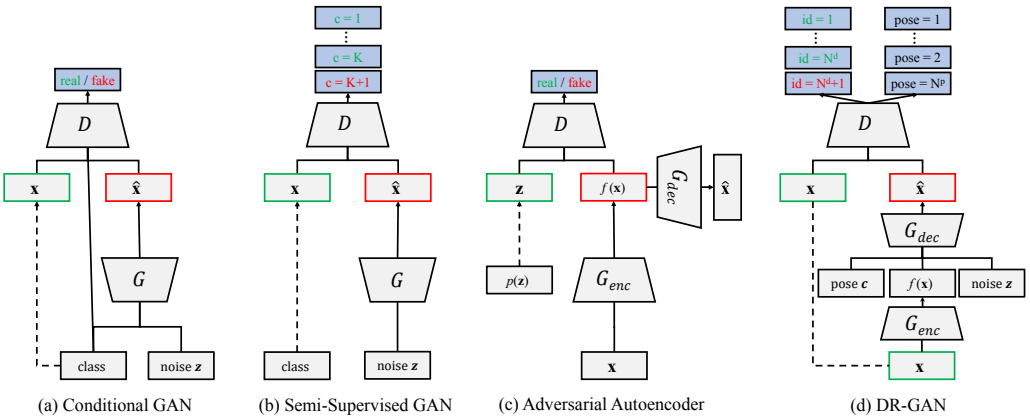
Generative Adversarial Network (GAN) Goodfellow et al. [9] introduce GAN to learn generative models via an adversarial process. With a minimax two-player game, the generator and discriminator can both improve themselves. GAN has been used for image synthesis [6,30], image super resolution [19,41], and etc. More recent work focus on incorporating constraints on $\mathbf{z}$ or leveraging side information for better synthesis. E.g., Mirza and Osindero [26] feed the class label to both G and D to generate images conditioned on the class label. Springenberg [35] generalizes GAN to learn a discriminative classifier where D is trained to not only distinguish between real and fake, but also classify the images. In InfoGAN [5], G applies information regularization to the optimization process, by using the additional latent code. In contrast, this paper proposes a novel DR-GAN for face *representation learning*, in addition to image synthesis. In Sec. 3.4, we will provide in-depth discussion on our difference to most relevant work in conventional GANs.

Face Frontalization Generating a frontal face from a profile face is very challenging due to self-occlusion. Existing methods for face frontalization can be classified into three categories: 3D-based methods [11,20,43], statistical methods [31], and deep learning methods [14,38,40,42,45]. E.g., Hassner et al. [11] use a mean 3D face model to generate a frontal face for any subject, which is proved to be effective and efficient. In [31], a statistical model is used for joint frontal view reconstruction and landmark localization by solving a constrained low-rank minimization problem. For deep learning methods, Kan et al. [14] propose SPAE to progressively rotate a non-frontal face image to a frontal face via auto-encoders. Yang et al. [38] apply the recurrent action unit to a group of hidden units to incrementally rotate faces in fixed yaw angles.

All prior work frontalize only near frontal in-the-wild faces [11,43] or large-pose controlled faces [40,45]. In contrast, we can synthesize arbitrary-pose faces from a large-pose in-the-wild face. We use the *adversarial loss* to improve the quality of synthetic images and identity classification in the discriminator to preserve the identity.

Representation Learning Designing the appropriate objectives for learning a good representation is an open question [1]. The work in [24] is among the first to use an encoder-decoder structure for representation learning, which, however, is not explicitly disentangled. DR-GAN is similar to DC-IGN [17] – a variational autoencoder-based method to disentangled representation learning. However, DC-IGN achieves disentanglement by providing batch training samples with one attribute being fixed, which may not be applicable to unstructured in-the-wild data.

Prior work also explore joint representation learning and face rotation for PIFR where [40, 45] are most relevant to our work. In [45], the authors propose Multi-View Perceptron [45] that can untangle the identity and view representations by processing them with different neurons and maximizing the data log-likelihood. Yim et al. [40] use a multi-task CNN to rotate a face with any pose and illumination to a target pose, and the $L2$ loss-based reconstruction of the input is the second task. Both work generate multi-view images and extract an identity representation. DR-GAN differs to [40,45] in two aspects. First, we explicitly disentangle the identity representation by using the pose code. Second, we employ the adversarial loss for high-quality synthesis, which drives better representation learning. Finally, none of them applies to in-the-wild faces as we do.



(a) Conditional GAN (b) Semi-Supervised GAN (c) Adversarial Autoencoder (d) DR-GAN

Figure 2: Comparison of previous GAN architectures and our proposed DR-GAN.

3. Proposed Method

DR-GAN has two variations: the basic model that takes one image as the input, termed as *single-image DR-GAN*, and the extended model that leverages multiple images per subject, termed as *multi-image DR-GAN*.

3.1. Generative Adversarial Network

Generative Adversarial Network (GAN) consists of a generator G and a discriminator D that compete in a two-player minimax game. D tries to distinguish a real image $\mathbf{x}$ from a synthetic one $G(\mathbf{z})$, and G tries to synthesize realistic-looking images that can fool D. Concretely, D and G play the game with a value function $V(D,G)$:

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad (1)$$

It is proved in [9] that this minimax game has a global optimum when the distribution p_g of the synthetic samples and the distribution p_d of the training samples are the same. Under mild conditions (e.g., G and D have enough capacity), p_g converges to p_d. In practice, it is better for G to maximize $\log(D(G(\mathbf{z})))$ instead of minimizing $\log(1 - D(G(\mathbf{z})))$ [9]. As a result, G and D are trained to alternatively optimize the following objectives:

$$\max_D V_D(D,G) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))], \quad (2)$$

$$\max_G V_G(D,G) = E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(D(G(\mathbf{z}))]. \quad (3)$$

3.2. Single-Image DR-GAN

The single-image DR-GAN has two distinctive novelties compared to prior GANs. First, it learns an identity representation for a face image by using an encoder-decoder structured generator, where the representation is the encoder's output and the decoder's input. Since the representation is the input to the decoder to synthesize various faces of the same subject, it is a *generative* representation.

Second, in face recognition, there are normally distractive variations existing in a face's appearance. Thus, the representation learned by the encoder might include the distractive side variation. E.g., the encoder would generate different identity representations for two faces of the same subject with $0°$ and $90°$ yaw. To remedy this, in addition to the class labels similar to semi-supervised GAN [35], we employ side information such as pose and illumination to explicitly disentangle these variations, which in turn helps to learn a *discriminative* representation.

3.2.1 Problem Formulation

Given a face image $\mathbf{x}$ with label $\mathbf{y} = \{y^d, y^p\}$, where y^d represents the label for identity and y^p for pose, the objectives of our learning problem are twofold: 1) to learn a pose-invariant identity representation for PIFR, and 2) to synthesize a face image $\hat{\mathbf{x}}$ with the *same* identity y^d but a *different* pose specified by a pose code $\mathbf{c}$. Our approach is to train a DR-GAN conditioned on the original image $\mathbf{x}$ and the pose code $\mathbf{c}$ with its architecture illustrated in Fig. 2 (d).

Different from the discriminator in conventional GAN, our D is a multi-task CNN consisting of two parts: $D = [D^d, D^p]$. $D^d \in \mathbb{R}^{N^d+1}$ is for identity classification with N^d as the total number of subjects in the training set and the additional dimension is for the fake class. $D^p \in \mathbb{R}^{N^p}$ is for pose classification with N^p as the total number of discrete poses. Given a real face image $\mathbf{x}$, D aims to estimate its identity and pose; while given a synthetic face image from the generator $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c}, \mathbf{z})$, D attempts to classify $\hat{\mathbf{x}}$ as fake, using the following objective:

$$\max_D V_D(D,G) = E_{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y})}[\log D^d_{y^d}(\mathbf{x}) + \log D^p_{y^p}(\mathbf{x})] + E_{\substack{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{N^d+1}(G(\mathbf{x},\mathbf{c},\mathbf{z})))], \quad (4)$$

where D^d_i and D^p_i are the ith element in D^d and D^p. The first term is to maximize the probability of $\mathbf{x}$ being classified to the true identity and pose. The second term is to maximize the probability of $\hat{\mathbf{x}}$ being classified as a fake class.

Table 1: The network structure of DR-GAN. Blue texts represent extra elements to learn the coefficient ω in multi-image DR-GAN.

G_{enc} and D			G_{dec}		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
			FC		$6 \times 6 \times 320$
Conv11	$3 \times 3/1$	$96 \times 96 \times 32$	FConv52	$3 \times 3/1$	$6 \times 6 \times 160$
Conv12	$3 \times 3/1$	$96 \times 96 \times 64$	FConv51	$3 \times 3/1$	$6 \times 6 \times 256$
Conv21	$3 \times 3/2$	$48 \times 48 \times 64$	FConv43	$3 \times 3/2$	$12 \times 12 \times 256$
Conv22	$3 \times 3/1$	$48 \times 48 \times 64$	FConv42	$3 \times 3/1$	$12 \times 12 \times 128$
Conv23	$3 \times 3/1$	$48 \times 48 \times 128$	FConv41	$3 \times 3/1$	$12 \times 12 \times 192$
Conv31	$3 \times 3/2$	$24 \times 24 \times 128$	FConv33	$3 \times 3/2$	$24 \times 24 \times 192$
Conv32	$3 \times 3/1$	$24 \times 24 \times 96$	FConv32	$3 \times 3/1$	$24 \times 24 \times 96$
Conv33	$3 \times 3/1$	$24 \times 24 \times 192$	FConv31	$3 \times 3/1$	$24 \times 24 \times 128$
Conv41	$3 \times 3/2$	$12 \times 12 \times 192$	FConv23	$3 \times 3/2$	$48 \times 48 \times 128$
Conv42	$3 \times 3/1$	$12 \times 12 \times 128$	FConv22	$3 \times 3/1$	$48 \times 48 \times 64$
Conv43	$3 \times 3/1$	$12 \times 12 \times 256$	FConv21	$3 \times 3/1$	$48 \times 48 \times 64$
Conv51	$3 \times 3/2$	$6 \times 6 \times 256$	FConv13	$3 \times 3/2$	$96 \times 96 \times 64$
Conv52	$3 \times 3/1$	$6 \times 6 \times 160$	FConv12	$3 \times 3/1$	$96 \times 96 \times 32$
Conv53	$3 \times 3/1$	$6 \times 6 \times (320$ +1)	FConv11	$3 \times 3/1$	$96 \times 96 \times 1$
AvgPool	$6 \times 6/1$	$1 \times 1 \times (320$ +1)			
FC (D only)		$N^d + N^p + 1$			

Meanwhile, G consists of an encoder G_{enc} and a decoder G_{dec}. G_{enc} aims to learn an identity representation from a face image $\mathbf{x}$: $f(\mathbf{x}) = G_{enc}(\mathbf{x})$. G_{dec} aims to synthesize a face image $\hat{\mathbf{x}} = G_{dec}(f(\mathbf{x}), \mathbf{c}, \mathbf{z})$ with identity y^d and a target pose specified by $\mathbf{c}$, where $\mathbf{z} \in \mathbb{R}^{N^z}$ is the noise modeling other variance besides identity or pose. The pose code $\mathbf{c} \in \mathbb{R}^{N^p}$ is a one-hot vector with the target pose y^t being 1. The goal of G is to fool D to classify $\hat{\mathbf{x}}$ to the identity of input $\mathbf{x}$ and the target pose with the following objective:

$$\max_G V_G(D,G) = E_{\substack{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x},\mathbf{c},\mathbf{z}))) + \log(D^p_{y^t}(G(\mathbf{x},\mathbf{c},\mathbf{z})))]. \quad (5)$$

G and D improves each other during alternative training. With D being more powerful in distinguishing real vs. fake images and classifying poses, G strives for synthesizing an identity-preserving face with the target pose to compete with D, with three benefits. First, the learnt representation $f(\mathbf{x})$ will preserve more discriminative identity information. Second, the pose classification in D guides the pose of the rotated face to be more accurate. Third, with a separate pose code input to G_{dec}, G_{enc} is trained to disentangle the pose variation from $f(\mathbf{x})$, i.e., $f(\mathbf{x})$ should encode as much identity information as possible, but as little pose information as possible. Thus, $f(\mathbf{x})$ is not only generative for image synthesis, but also discriminative for PIFR.

3.2.2 Network Structure

The network structure of single-image DR-GAN is shown in Tab. 1. We adopt CASIA-Net [39] for G_{enc} and D where batch normalization (BN) and exponential linear unit (ELU) are applied after each convolutional layer. D is trained to optimize Eqn. 4 by adding a fully connected layer with softmax loss for $(N_d + 1)$ identity and (N_p) pose classification. G includes G_{enc} and G_{dec} that are bridged by the to-be-learned identity representation $f(\mathbf{x}) \in \mathbb{R}^{320}$, which is the AvgPool output in our network. $f(\mathbf{x})$ is concatenated

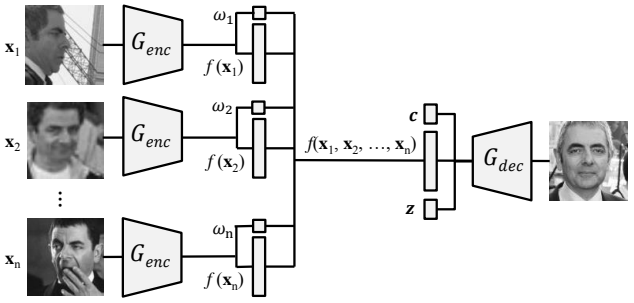


Figure 3: Generator in multi-image DR-GAN. From an image set of a subject, we can fuse the features to a single representation via dynamically learnt coefficients and synthesize images in any pose.

with a pose code $\mathbf{c}$ and a random noise $\mathbf{z}$. A series of fractionally-strided convolutions (FConv) [29] transforms the $(320 + N^p + N^z)$-dim concatenated vector into a synthetic image $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c}, \mathbf{z})$, which is the same size as $\mathbf{x}$. G is trained to maximize Eqn. 5 when a synthetic face $\hat{\mathbf{x}}$ is fed to D and the gradient is back-propagated to update G.

3.3. Multi-Image DR-GAN

Single-Image DR-GAN extracts an identity representation and performs face rotation by processing one single image $\mathbf{x}$. Yet, we often have multiple images per subject in training and sometimes in testing. To leverage them, we propose multi-image DR-GAN that can benefit both the training and testing stages. For training, it can learn a better identity representation from multiple images that are complementary to each other. For testing, it can enable template-to-template matching, which addresses a crucial need in real-world surveillance applications.

Multi-Image DR-GAN has the same D as single-image DR-GAN, but a different G as shown in Fig. 3. Besides extracting $f(\mathbf{x})$, G_{enc} also estimates a confident coefficient ω for each image, which predicts the quality of the learnt representation. With n input images $\{\mathbf{x}_i\}_{i=1}^n$, the fused representation is the weighted average of all representations,

$$f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) = \frac{\sum_{i=1}^n \omega_i f(\mathbf{x}_i)}{\sum_{i=1}^n \omega_i}. \quad (6)$$

The fused representation is concatenated with $\mathbf{c}$ and $\mathbf{z}$ and fed to G_{dec} to generate a new image, which is expected to have the same identity as all input images. Thus, the objective for learning G has a total of $2(n + 1)$ terms:

$$\begin{aligned}\max_G V_G(D,G) = &\sum_{i=1}^n [E_{\substack{\mathbf{x}_i,\mathbf{y}_i\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x}_i,\mathbf{c},\mathbf{z}))) + \\ &\log(D^p_{y^t}(G(\mathbf{x}_i,\mathbf{c},\mathbf{z})))]] + \\ &E_{\substack{\mathbf{x}_i,\mathbf{y}_i\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x}_1,...,\mathbf{x}_n,\mathbf{c},\mathbf{z}))) + \\ &\log(D^p_{y^t}(G(\mathbf{x}_1,...,\mathbf{x}_n,\mathbf{c},\mathbf{z})))]. \quad (7)\end{aligned}$$

The coefficient ω_i is learned so that an image with a higher quality contributes more to the fused representation. Here the quality can be viewed as an indicator of the PIFR

performance, rather than the low-level image quality. Face quality prediction is a classic topic where many prior work attempt to estimate the former from the latter [27, 37]. Our coefficient learning is essentially quality prediction, from novel perspectives in contrast to prior work. That is, without explicit supervision, it is driven by D through the decoded image $G_{dec}(f(\mathbf{x}_1, ..., \mathbf{x}_n), \mathbf{c}, \mathbf{z})$, and learned in the context of, as a byproduct of, representation learning.

Note that, jointly training multiple images per subject results in *one*, but not multiple, generator, i.e., all G_{enc} in Fig. 3 share the same parameters. This makes it flexible to take an *arbitrary number* of images during testing for representation learning via Eqn. 6 and face rotation. While the network in Fig. 2 (d) is used for training, our network for testing is much simplified: only G_{enc} is used to extract representations; both G_{enc} and G_{dec} are used for face rotation.

For the network structure, multi-image DR-GAN only makes minor modification from the single-image counterpart. Specifically, at the end of G_{enc}, we add one more convolution channel to the layer before AvgPool, to estimate the coefficient ω. We apply $Sigmoid$ activation to constrain ω in the range of $[0, 1]$. During training, despite unnecessary, we keep the number of input images per subject n the same for the sake of convenience in image sampling and network training. To mimic the variation in the number of input images, we use a simple but effective trick: applying Dropout on the coefficients ω. Hence, during training, the network takes any number of inputs varying from 1 to n.

3.4. Comparison to Prior GANs

We compare DR-GAN with three most relevant GAN variants, as shown in Fig 2.

Conditional GAN Conditional GAN [18, 26] extends the GAN by feeding the labels to both G and D to generate images conditioned on the label, which can be the class label, modality information, or even partial data for inpainting. It has been used to generate MNIST digits conditioned on the class label and to learn multi-modal models. In conditional GAN, D is trained to classify a real image with mismatched conditions to a fake class. In DR-GAN, D classifies a real image to the corresponding class based on the labels.

Semi-Supervised GAN Springenberg [35] generalizes GAN to learn a discriminative classifier where D is trained to not only distinguish between real and fake images, but also classify real images into K classes. D outputs a $(K+1)$-dim vector with the last dimension for the real/fake decision. The trained D is used for image classification. DR-GAN shares a similar loss for D as [35] but has two additions. First, we expand G with an encoder-decoder structure. Second, we have an additional side information classification on the pose while training D.

Adversarial Autoencoder (AAE) In AAE [23], G is the encoder of an autoencoder. AAE has two objectives in order to turn an autoencoder into a generative model: the autoencoder reconstructs the input image, and the latent vector generated by the encoder matches an arbitrary prior distribution by training D. DR-GAN differs to AAE in two aspects. First, the autoencoder in [23] is trained to learn a latent representation similar to an imposed prior distribution, while our encoder-decoder learns discriminative identity representations. Second, D in AAE is trained to distinguish real/fake distributions while our D is trained to classify real/fake images, the identity and pose of the images.

4. Experimental Results

DR-GAN aims for both representation learning and face synthesis. For the former, we quantitatively evaluate the face recognition performance using the disentangled representation as the identity features with a cosine distance metric, for both the controlled and in-the-wild settings. For the latter, we show qualitative results of face frontalization.

4.1. Experimental Settings

Databases Multi-PIE [10] is the largest database for evaluating face recognition under pose, illumination, and expression variations in the controlled setting. Following the setting in [45], we use 337 subjects with neutral expression, 9 poses within $\pm 60^o$, and 20 illuminations. The first 200 subjects are for training and the rest 137 for testing. For testing, one image per subject with the frontal view and neutral illumination is the gallery and the others are the probes. For Multi-PIE experiments, we add an additional illumination code similar to the pose code to disentangle the illumination. Therefore, we have $N^d = 200$, $N^p = 9$, $N^{il} = 20$.

For the in-the-wild setting, we train on Multi-PIE and CASIA-WebFace [39], and test on CFP [34] and IJB-A [16]. CASIA-WebFace includes $494,414$ near-frontal faces of $10,575$ subjects. We add the entire Multi-PIE (4 sessions, 13 poses, 6 expressions, and 20 illuminations of 337 subjects) to the training set to supply more pose variation. CFP consists of 500 subjects each with 10 frontal and 4 profile images. The evaluation protocol includes frontal-frontal (FF) and frontal-profile (FP) face verification, each having 10 folders with 350 same-person pairs and 350 different-person pairs. As another large-pose database, IJB-A has $5,396$ images and $20,412$ video frames of 500 subjects. It defines template-to-template face recognition where each template has one or multiple images. We remove 27 overlap subjects between CASIA-Webface and IJB-A from the training. We have $N^d = 10,885$, $N^p = 13$. We set $N^z = 50$ for both settings.

Implementation Details Following [39], we align all face images to a canonical view of size 100×100. We randomly sample 96×96 regions from the aligned 100×100 face images for data augmentation. Image intensities are linearly scaled to the range of $[-1, 1]$. To provide pose labels y^p for CASIA-WebFace, we apply 3D face alignment [12, 13]

Table 2: Performance comparison on CFP.

Method	Frontal-Frontal	Frontal-Profile
Sengupta et al. [34]	96.40 ± 0.69	84.91 ± 1.82
Sankarana et al. [32]	96.93 ± 0.61	89.17 ± 2.35
Chen et al. [4]	$\mathbf{98.67} \pm 0.36$	91.97 ± 1.70
Human	96.24 ± 0.67	94.57 ± 1.10
DR-GAN: synthetic	97.08 ± 0.62	91.02 ± 1.59
DR-GAN: n=1	97.13 ± 0.68	90.82 ± 0.28
DR-GAN: n=4	97.86 ± 0.75	92.93 ± 1.39
DR-GAN: n=6	97.84 ± 0.79	$\mathbf{93.41} \pm 1.17$

to classify each face to one of 13 poses. Our implementation is extensively modified from a publicly available implementation of DC-GAN. We follow the optimization strategy in [29]. The batch size is set to be 64. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.02. Adam optimizer [15] is used with a learning rate of 0.0002 and momentum 0.5.

In conventional GANs, Goodfellow et al. [9] suggest to alternate between k (usually $k = 1$) steps of optimizing D and one step of optimizing G. This helps D to maintain near-optimal solution as long as G changes slowly. However, in DR-GAN, D has strong supervisions thanks to the class labels. Thus, in later iterations, when D is close to the optimal solution, we update G more frequently than D, e.g., 4 steps for optimizing G and 1 for D.

4.2. Representation Learning

Single vs. Multiple Training Images We evaluate the effect of the number of training images (n) per subject on the face recognition performance. Specifically, with the same training set, we train three models with $n = 1, 4, 6$, where $n = 1$ denotes single-image DR-GAN and $n > 1$ denotes multi-image DR-GAN. The testing performance on CFP using $f(\mathbf{x})$ of each model are shown in Tab. 2. We observe the advantage of multi-image DR-GAN over the single-image counterpart, which attributes to more constraints in learning G_{enc} that leads to a better representation. However, we do not keep increasing n due to the limited computation capacity. In the rest of the paper, we use multi-image DR-GAN with $n = 6$ unless specified.

Single vs. Multiple Testing Images We also evaluate the effect of the number of testing images (n_t) per subject on the face recognition performance on Multi-PIE. We mimic IJB-A to generate image sets as the probe set while the gallery set remains the same with one image per subject. Specifically, from the Multi-PIE probe set, we select a subsect $\mathbb{P}_0$ of images with large poses (30° to 60°), which are used to form 5 different probe sets $\{\mathbb{P}_i\}_{i=1}^5$ with n_t ranging from 1 to 5. First, we randomly select one image per subject from $\mathbb{P}_0$ to form $\mathbb{P}_1$. Second, based on $\mathbb{P}_1$, we construct $\mathbb{P}_2$ by adding one random image of each subject from $\mathbb{P}_0$. We construct $\mathbb{P}_3$, $\mathbb{P}_4$, $\mathbb{P}_5$ in a similar way.

We compare three combinations of models and decision metrics: (i) single-image DR-GAN with the averaged cosine distances of n_t representations, (ii) multi-image DR-GAN with the averaged cosine distances of n_t representations, and (iii) multi-image DR-GAN with the cosine distance of the fused representation. As shown in Tab. 3, comparing (ii) and (iii), using the coefficients learned by the network for representation fusion is superior over the conventional score averaging. There is a consistent improvement of $\sim$0.5%. While there is some improvement from (i) to (ii), the margin decreases as n_t increases.

Table 3: Identification rates of three approaches on Multi-PIE.

n_t	1	2	3	4	5
single-image (avg.)	84.6	91.8	94.1	95.3	95.8
multi-image (avg.)	85.9	92.4	94.5	95.5	95.9
multi-image (fuse)	85.9	92.8	95.1	96.0	96.5

Table 4: Benchmark comparison on Multi-PIE.

Method	0°	15°	30°	45°	60°	Average
Zhu et al. [44]	94.3	90.7	80.7	64.1	45.9	72.9
Zhu et al. [45]	95.7	92.8	83.7	72.9	60.1	79.3
Yim et al. [40]	**99.5**	**95.0**	88.5	79.9	61.9	83.3
Using $L2$ loss	95.1	90.8	82.7	72.7	57.9	78.3
DR-GAN	97.0	94.0	**90.1**	**86.2**	**83.2**	**89.2**

Results on Benchmark Databases We compare our method with state-of-the-art face recognition methods on Multi-PIE, CFP, and IJB-A. Table 4 shows the face identification performance on Multi-PIE compared to the methods with the same setting. Our method shows a significant improvement for large-pose faces. The variation of recognition rates across different poses is much smaller than the baselines, which suggests that our learnt representation is more robust to pose variation.

Table 2 shows the comparison on CFP. Results are reported with the average face verification accuracy with standard deviation over 10 folds. We achieve comparable performance on frontal-frontal verification while having $\sim$1.4% improvement on the frontal-profile verification.

Table 5 shows the performance of both face identification and verification on IJB-A. DR-GAN achieves comparable performance to state-of-the-art methods. Further, the proposed fusion scheme via learnt coefficients is superior to the averaged cosine distances of representations. These in-the-wild results show the power of DR-GAN for PIFR.

4.3. Face Synthesis

Adversarial Loss vs. $L2$ Loss Prior work [38, 40, 44] on face rotation normally employ the $L2$ loss to learn a mapping between two views. To compare the $L2$ loss with our adversarial loss, we train a model where G is supervised by an $L2$ loss on the ground truth face with the target view. The training process is kept the same for a fair comparison. As shown in Fig. 4, DR-GAN can generate far more real-

Table 5: Performance comparison on IJB-A.

Method	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
OpenBR [16]	23.6 ± 0.9	10.4 ± 1.4	24.6 ± 1.1	37.5 ± 0.8
GOTS [16]	40.6 ± 1.4	19.8 ± 0.8	44.3 ± 2.1	59.5 ± 2.0
Wang et al. [36]	72.9 ± 3.5	51.0 ± 6.1	82.2 ± 2.3	93.1 ± 1.4
PAM [25]	73.3 ± 1.8	$\mathbf{55.2} \pm 3.2$	77.1 ± 1.6	88.7 ± 0.9
DCNN [3]	$\mathbf{78.7} \pm 4.3$	–	85.2 ± 1.8	93.7 ± 1.0
DR-GAN (avg.)	75.5 ± 2.8	51.8 ± 6.8	84.3 ± 1.3	93.2 ± 0.8
DR-GAN (fuse)	77.4 ± 2.7	53.9 ± 4.3	$\mathbf{85.5} \pm 1.5$	$\mathbf{94.7} \pm 1.1$

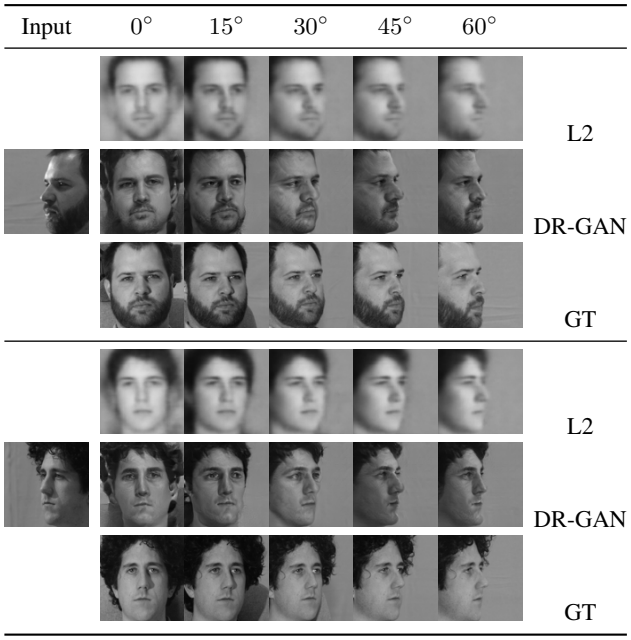


Figure 4: Visual comparison of face synthesis on Multi-PIE. For each input image, we compare synthetic images of $L2$ loss (top), adversarial loss (middle), and their ground truth (bottom).

istic faces that are similar to the ground truth faces in all views. Meanwhile, images synthesized by the $L2$ loss cannot maintain high frequency components and are blurry. In fact, $L2$ loss treats each pixel equally, which leads to the loss of discriminative information. This inferior synthesis is also reflected in the lower PIFR performance in Tab. 4.

Interpolation of Variables Taking two images of different subjects $\mathbf{x}_1, \mathbf{x}_2$, we extract features $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$ from G_{enc}. The interpolation between $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$ can generate many representations, which can be fed to G_{dec} to synthesize face images. In Fig. 5 (a), the top row shows a transition from a male subject with beard and glasses to a female without these adjectives. Similar to [29], these smooth semantic changes indicate that the model has learned essential identity representations for image synthesis.

During training, we use a one-hot vector $\mathbf{c}$ to specify the *discrete* pose of the synthetic image. During testing, we could also interpolate between two neighboring pose codes, to generate face images with *continuous* poses. As in Fig. 5 (b), this leads to smooth pose transition from one view to many views *unseen* to the training set.

We also interpolate the noise $\mathbf{z}$. We synthesize frontal faces at $\mathbf{z} = -\mathbf{1}$ and $\mathbf{z} = \mathbf{1}$ (a vector of all 1s) and interpolate between two $\mathbf{z}$. Given the fixed identity representation and pose code, the synthetic images are identity-preserving frontal faces. For better visualization, we show the difference of the images w.r.t. the image generated by $\mathbf{z} = -\mathbf{1}$. As in Fig. 5 (c), $\mathbf{z}$ models less significant face variations.

Face Rotation Our generator is trained to be a face rotator. Face rotation on Multi-PIE is shown in Fig. 4 and Fig. 5. Figure 6 shows the face frontalization on CFP. Given an input image at large poses even extreme profile, DR-GAN can generate a realistic frontal face that is very similar to the real frontal face. To the best of our knowledge, this is the first work that is able to *frontalizate a profile-view in-the-wild face image*. Figure 7 shows the face rotation on IJB-A with various number of input images. During face rotation, the identity is preserved and pose-view is changed exactly to the target pose, indicating that the learnt representation is largely disentangled with other variations and the pose code entirely determines the synthesized pose. We want to emphasize that it is extremely difficult to *fuse multiple in-the-wild face images into a single frontal face image*, when the input images have diverse and large variations in poses, expressions, lightings and resolutions. We believe that this is the first time such a fusing capability has been demonstrated on challenging databases such as IJB-A.

Representation vs. Synthetic Image DR-GAN can simultaneously extract a representation, and generate a frontal face of the same subject - both are useful for PIFR. The representation is directly used via the cosine distance metric. The synthetic image can be fed to a pretrained face recognition model, with an architecture similar to CASIA-Net [39] to extract identity features for PIFR. The performance comparison is reported in Tab. 2. It appears that the representation is more effective. However, the synthetic images are visually appealing, and can be promising in law enforcement practice, especially given their notable PIFR performance.

5. Conclusions

This paper presents DR-GAN for pose-invariant face recognition and face synthesis. We extend GAN with a few distinct novelties, including the encoder-decoder structured generator, pose code, pose classification in the discriminator, and an integrated multi-image fusion scheme. We attribute the superior PIFR performance and face synthesis capability to the discriminative yet generative representation learned in the generator. Our learnt representation is discriminative because the other variations have been explicitly disentangled by the pose code and the illumination code, and is generative because its decoded (synthetic) image would still be classified as the original identity.

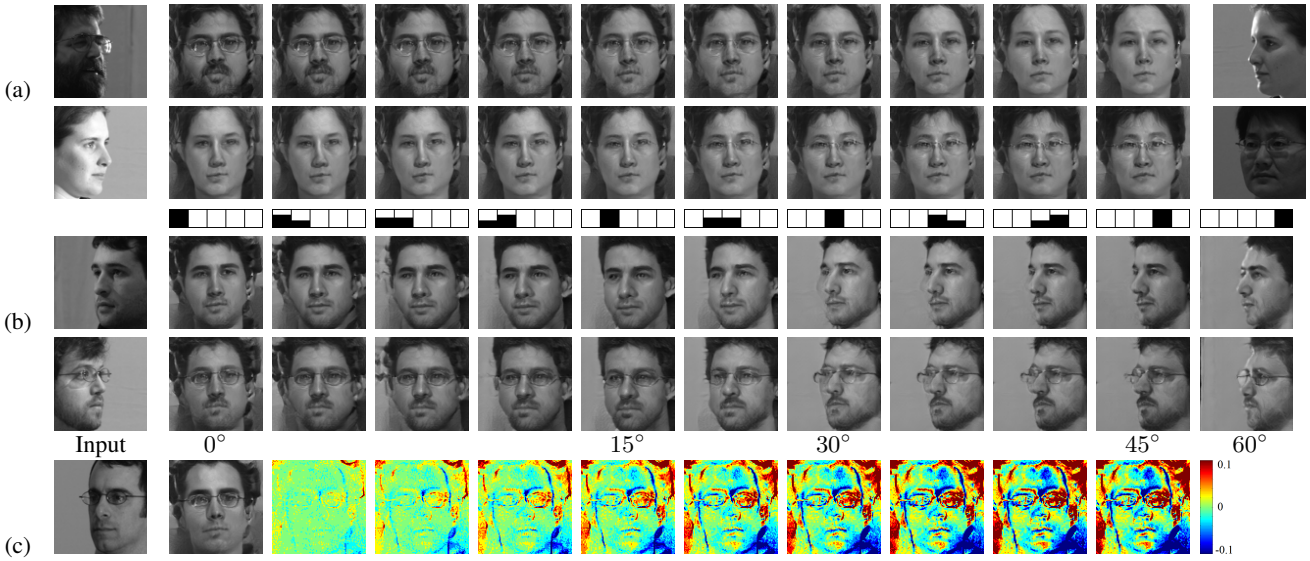


Figure 5: (a) Image synthesis by interpolating between the identity representations of two faces (far left and right). (b) While only discrete poses are available in training, DR-GAN can synthesize new poses by interpolating *continuous* pose codes, shown above Row 3. (c) The input image, rotated image of $\mathbf{z} = -\mathbf{1}$, and its differences to 9 in-between images toward the rotated image of $\mathbf{z} = \mathbf{1}$.

Figure 6: Face frontalization on CFP. From top to bottom: input images, our frontalized faces, real frontal faces. We only expect the frontalized faces to be similar to real faces in the identity, rather than in all facial attributes. This is very challenging for face rotation due to the in-the-wild variations and extreme profile views. The artifact in the image boundary is due to image extrapolation in pre-processing.

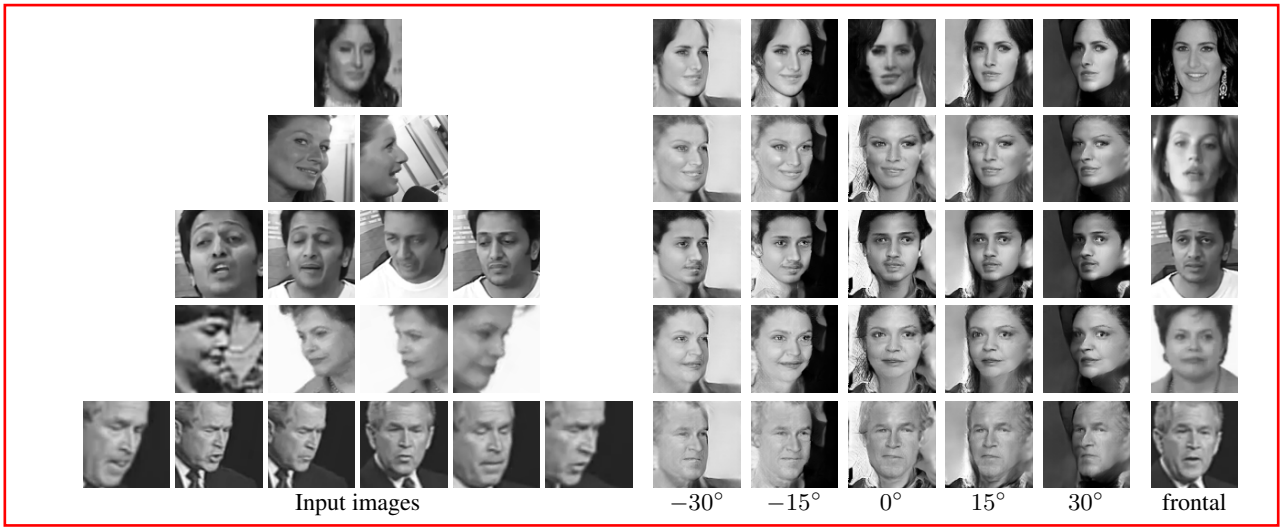


Figure 7: Multi-image face rotation on IJB-A. For each subject, we show $1-6$ input images, synthetic images at 5 poses, and a real frontal face. In additional to the profile view in CFP, this task has more challenges: 1) the low image quality of the input; 2) the large variations within a set, such as poses, resolution, and expression. DR-GAN seems also super-resolving faces and neutralizing expressions.