

Representation Learning by Rotating Your Faces

Luan Tran, Xi Yin, and Xiaoming Liu, *Member, IEEE*

Abstract—The large pose discrepancy between two face images is one of the fundamental challenges in automatic face recognition. Conventional approaches to 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 a Disentangled Representation learning-Generative Adversarial Network (DR-GAN) with three distinct novelties. First, the encoder-decoder structure of the generator enables DR-GAN to learn a representation that is both generative and discriminative, which can be used for face image synthesis and pose-invariant face recognition. 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 identity representation along with an arbitrary number of synthetic face images. Extensive quantitative and qualitative evaluation on a number of controlled and in-the-wild databases demonstrate the superiority of DR-GAN over the state of the art in both learning representations and rotating large-pose face images.

Index Terms—representation learning, generative adversarial network, pose-invariant face recognition, face rotation and frontalization

1 INTRODUCTION

FACE recognition is one of the most widely studied topics in computer vision due to its wide application in law enforcement, biometrics, marketing, and etc. Recently, great progress has been achieved in face recognition with deep learning-based methods [1], [2], [3]. For example, surpassing human performance is reported by Schroff et al. [3] on Labeled Faces in the Wild (LFW) database. However, one of the shortcomings of the LFW database is that it does not offer a high degree of pose variation — the variance that has been shown to be a major challenge in face recognition. Up to now, the key ability of Pose-Invariant Face Recognition (PIFR) desired by real-world applications is far from solved [4], [5], [6], [7], [8]. A recent study [9] observes a significant drop, over 10%, in performance of most algorithms from frontal-frontal to frontal-profile face verification, while human performance only degrades slightly. This indicates that the pose variation remains to be a significant challenge in face recognition and warrants future study.

In PIFR, the facial appearance change caused by pose variation often significantly surpasses the intrinsic appearance differences between individuals. To overcome these challenges, a wide variety of approaches have been proposed, which can be grouped into two categories. First, some work employ *face frontalization* on the input image to synthesize a frontal-view face, where traditional face recognition algorithms are applicable [10], [11], or an identity representation can be obtained via modeling the face frontalization/rotation process [12], [13], [14]. The ability to generate a realistic identity-preserved frontal face is also beneficial for law enforcement practitioners to identify suspects. Second, other work focus on *learning discriminative representations* directly from the non-frontal faces through either one joint model [2], [3] or multiple pose-specific models [15], [16]. In contrast, we propose a novel framework to take the best of both worlds — *simultaneously learn pose-invariant identity representation and synthesize faces*

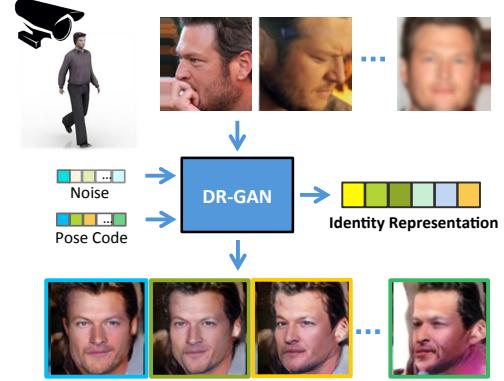


Fig. 1: Given one or multiple in-the-wild face images as the input, DR-GAN can produce a unified identity representation, by virtually rotating the face to arbitrary poses. The learnt representation is both *discriminative* and *generative*, i.e., the representation is able to demonstrate superior PIFR performance, and synthesize identity-preserved faces at target poses specified by the pose code.

with *arbitrary poses*, where face rotation is both a facilitator and a by-product for representation learning.

As shown in Fig. 1, we propose Disentangled Representation learning-Generative Adversarial Network (DR-GAN) for PIFR. Generative Adversarial Networks (GANs) [17] can generate samples following a data distribution through a two-player game between a generator G and a discriminator D . Despite many recent promising developments [18], [19], [20], [21], [22], image synthesis remains to be the main objective of GAN. To the best of our knowledge, this is the first work that utilizes the generator in GAN for representation learning. To achieve this, we conduct G with an encoder-decoder structure (Fig. 2 (d)) to learn a disentangled representation for PIFR. 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 (or fake) images,

• L. Tran, X. Yin and X. Liu are with the Department of Computer Science and Engineering, Michigan State University. E-mail: tranluan@msu.edu, yinxil@msu.edu, liuxm@cse.msu.edu

Manuscript received May 29, 2017.

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-preserved 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 c , and a random noise vector z as the input, with the objective 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. Note that it is a crucial architecture design to concatenate one representation with *varying* randomly generated pose codes and noise vectors. This enables DR-GAN to learn a *disentangled* identity representation that is *exclusive* or *invariant* to pose and other variations, which is the holy grail for PIFR when achievable.

Most existing face recognition algorithms only takes one image for testing. In practice, there are many scenarios when an image collection of the same individual is available [23]. In this case, prior work fuse results either in the feature level [24] or the distance-metric level [25], [26]. Differently, our fusion is conducted within a unified framework. Given multiple images as the input, G_{enc} operates on each image, and produces an identity representation and a coefficient, which is an indicator of the quality of that input image. Using the dynamically learned coefficients, the representations of all input images are linearly combined as one representation. During testing, G_{enc} takes any number of images and generates a single identity representation, which is used by G_{dec} for face synthesis along with the pose code.

Our generator is essential to both representation learning and image synthesis. We propose two techniques to further improve G_{enc} and G_{dec} respectively. First, we have observed that our G_{enc} can always outperform D in representation learning for PIFR. Therefore, we propose to replace the identity classification part of D with the latest G_{enc} during training so that a superior D can push G_{enc} to further improve itself. Second, since our G_{dec} learns a mapping from the feature space to the image space, we propose to improve the learning of G_{dec} by regularizing the average representation of two representations from different subjects to be a valid face, assuming a convex space of face identities. These two techniques are shown to be effective in improving the generalization ability of DR-GAN.

A preliminary version of this work was published in 2017 IEEE Conference on Computer Vision and Pattern Recognition [27]. We extend it in numerous ways: 1) Instead of having an extra dimension of the fake class in the identity classification task of the discriminator, we split it into two tasks: real/fake and identity classification. 2) We propose two techniques to improve model generalization during training. 3) We conduct all experiments using the new models with color image input, and add numerous experiments to reveal how DR-GAN works including the disentangled representation, the coefficients analysis, etc.

In summary, this paper makes the following contributions.

- 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.

- Our learnt representation is explicitly disentangled from the pose variation via the pose code in the generator and the pose estimation in the discriminator. Similar disentanglement is conducted for other variations, e.g., illumination.
- We propose a novel scheme to adaptively fuse multiple faces to a single representation based on the learnt coefficients, which empirically shows to be a good indicator of the face image quality.
- We propose two techniques to improve the generalization ability of our generator via model switch and representation interpolation.
- We achieve state-of-the-art face frontalization and face recognition performance on multiple benchmark datasets, including Multi-PIE [28], CFP [9], and IJB-A [23].

2 PRIOR WORK

Generative Adversarial Network (GAN). Goodfellow *et al.* [17] 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 [19], [29], image super resolution [30], and etc. More recent work focus on incorporating constraints to z or leveraging side information for better synthesis. E.g., Mirza and Osindero [18] feed class labels to both G and D to generate images conditioned on class labels. In [31] and [32], GAN is generalized to learn a discriminative classifier where D is trained to not only distinguish between real vs. fake, but also classify the images. In InfoGAN [21], G applies information regularization to the optimization by using the additional latent code. In contrast, this paper proposes a novel DR-GAN aiming for face *representation learning*, which is achieved via modeling the face rotation process. In Sec. 3.6, we will provide in-depth discussion on our difference to most relevant work in GANs.

One crucial issue with GANs is the difficulty for quantitative evaluation. Previous work either perform human study to evaluate the quality of synthetic images [19] or use the features in the discriminator for image classification [20]. In contrast, we innovatively construct the generator for representation learning, which can be quantitatively evaluated for PIFR.

Face Frontalization. Generating a frontal face from a profile face is very challenging due to self-occlusion. Prior methods in face frontalization can be classified into three categories: 3D-based methods [10], [11], [33], statistical methods [34], and deep learning methods [12], [13], [14], [35], [36]. E.g., Hassner *et al.* [10] use a mean 3D face model to generate a frontal face for any subject. A personalized face model could be used but accurate 3D face reconstruction remains a challenge [37], [38]. In [34], a statistical model is used for joint frontalization and landmark localization by solving a constrained low-rank minimization problem. For deep learning methods, Kan *et al.* [12] propose SPAE to progressively rotate a non-frontal face to a frontal one via auto-encoders. Yang *et al.* [35] 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 [10], [11] or large-pose controlled faces [13], [14]. 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 the synthetic images and identity classification in the discriminator to preserve identity.

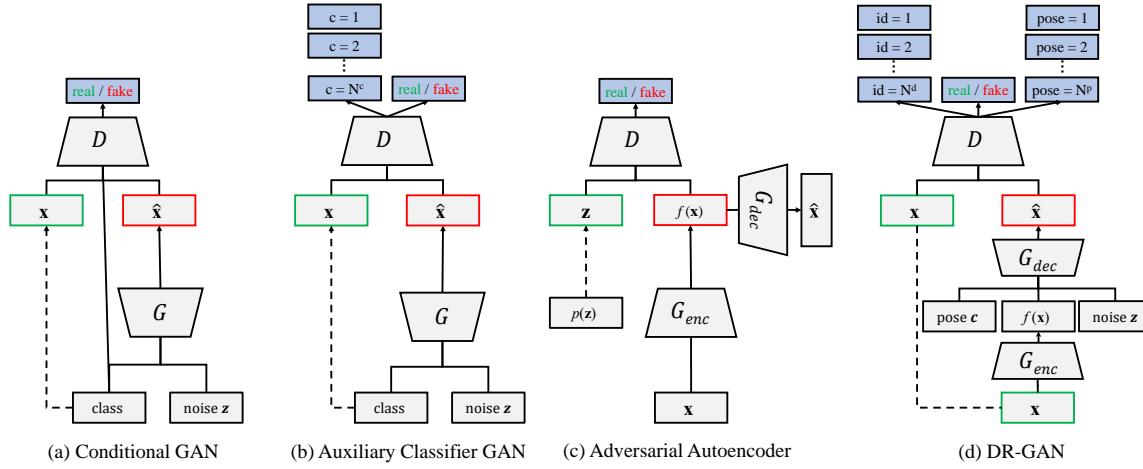


Fig. 2: Comparison of previous GAN architectures and our proposed DR-GAN.

Representation Learning. Designing the appropriate objectives for learning a good representation is an open question [39]. The work in [40] 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 [41] — 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 [13], [14] are most relevant to our work. In [13], Multi-View Perceptron [13] is used to untangle the identity and view representations by processing them with different neurons and maximizing the data log-likelihood. Yim *et al.* [14] use a multi-task CNN to rotate a face with any pose and illumination to a target pose, and the L_2 loss-based reconstruction of the input is the second task. Both work focus on image synthesis and the identity representation is a by-product during the network learning. In contrast, DR-GAN focuses on representation learning, of which face rotation is both a facilitator and a by-product. We differ to [13], [14] in four aspects. First, we explicitly disentangle the identity representation from pose variations by pose codes. Second, we employ the adversarial loss for high-quality synthesis, which drives better representation learning. Third, none of them applies to in-the-wild faces as we do. Finally, our ability to learn the representation from multiple unconstrained images has not been observed in prior work.

Face Image Quality Estimation. Image quality estimation is important for biometric recognition systems [42], [43], [44]. Numerous methods have been proposed to measure the image quality of different biometric modalities including face [45], [46], [47], iris [48], [49], fingerprint [50], [51], and gait [52], [53]. In the scenario of face recognition, an effective algorithm for face image quality estimation can help to either (i) reduce the number of poor images acquired during enrollment, or (ii) improve feature fusion during testing. Both cases can improve the face recognition performance. Abaza *et al.* [45] evaluate multiple quality factors such as contrast, brightness, sharpness, focus and illumination as a face image quality index for face recognition. However, they did not consider pose variance, which is a major challenge in face recognition. Ozay *et al.* [47] employ a Bayesian network to model the relationships between predefined quality related image features

and face recognition, which is show to boost the performance significantly. The authors in [54] propose a patch-based face image quality estimation method, which takes into account of geometric alignment, pose, sharpness, and shadows.

In this work, we employ quality estimation in a unified GAN framework that considers all factors of image quality presented in the dataset, with *no* direct supervision. For each input image, DR-GAN can generate a coefficient that indicates the quality of the input image. The representations from multiple images of the same subject are fused based on the learnt coefficients to generate one unified representation. We will show that the learnt coefficients are correlated to the image quality, i.e., a measurement of how good it can be used for face recognition.

3 THE PROPOSED DR-GAN MODEL

Our proposed DR-GAN has two variations: the basic model can take one image per subject for training, termed *single-image DR-GAN*, and the extended model can leverage multiple images per subject for both training and testing, termed *multi-image DR-GAN*. We start by introducing the original GAN, followed by two DR-GAN variations, and the proposed techniques to improve the generalization of our generator. Finally, we will compare our DR-GAN with previous GAN variations in detail.

3.1 Generative Adversarial Network

Generative Adversarial Network consists of a generator G and a discriminator D that compete in a two-player minimax game. The discriminator D tries to distinguish between a real image \mathbf{x} and a synthetic image $G(\mathbf{z})$. The generator G tries to synthesize realistic-looking images from a random noise vector \mathbf{z} that can fool D , i.e., $G(\mathbf{z})$ being classified as a real image. Concretely, D and G play the game with the following loss function:

$$\min_G \max_D \mathcal{L}_{gan} = \mathbb{E}_{\mathbf{x} \sim p_d(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (1)$$

It is proved in [17] that this minimax game has a global optimum when the distribution p_g of the synthetic samples and the distribution p_d of the real samples are the same. Under mild conditions (e.g., G and D have enough capacity), p_g converges to p_d . In the beginning of training, the samples generated from G are extremely poor and are rejected by D with high confidences.

In practice, it is better for G to maximize $\log(D(G(\mathbf{z})))$ instead of minimizing $\log(1 - D(G(\mathbf{z})))$ [17]. This objective results in the same fixed point of the dynamics of G and D but provides much stronger gradients early in learning. As a result, G and D are trained to alternatively optimize the following objectives:

$$\max_D \mathcal{L}_{gan}^D = \mathbb{E}_{\mathbf{x} \sim p_d(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))], \quad (2)$$

$$\max_G \mathcal{L}_{gan}^G = \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(D(G(\mathbf{z})))]. \quad (3)$$

3.2 Single-Image DR-GAN

Our 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, i.e., virtually rotating his/her face, it is a *generative* representation.

Second, the appearance of a face is determined by not only the identity, but also the numerous distractive variations, such as pose, illumination, expression. Thus, the identity representation learned by the encoder would inevitably include the distractive side variations. E.g., the encoder would generate *different* identity representations for two faces of the same subject with 0° and 90° yaw angles. To remedy this, in addition to the class labels similar to semi-supervised GAN [31], 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 at 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 three components: $D = [D^r, D^d, D^p]$. $D^r \in \mathbb{R}^1$ is for real/fake image classification. $D^d \in \mathbb{R}^{N^d}$ is for identity classification with N^d as the total number of subjects in the training set. $D^p \in \mathbb{R}^{N^p}$ is for pose classification with N^p as the total number of discrete poses. Note that, in our preliminary work [27], D^r is implemented as an additional $N^d + 1^{\text{th}}$ element of D^d , which has the problem of unbalanced training data for each dimension in D^d , i.e., the number of synthetic images ($N^d + 1^{\text{th}}$ dimension) equals to the summation of all images in the real classes (the first N^d dimensions). This version fixes this problem and is referred as “split” in Tab. 8. Given a face image \mathbf{x} , D aims to classify it as the real image class, and 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 objectives:

$$\mathcal{L}_{gan}^D = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_d(\mathbf{x}, \mathbf{y})} [\log D^r(\mathbf{x})] + \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_d(\mathbf{x}, \mathbf{y}), \mathbf{z} \sim p_z(\mathbf{z}), \mathbf{c} \sim p_c(\mathbf{c})} [\log(1 - D^r(G(\mathbf{x}, \mathbf{c}, \mathbf{z})))], \quad (4)$$

$$\mathcal{L}_{id}^D = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_d(\mathbf{x}, \mathbf{y})} [\log D_{y^d}^d(\mathbf{x})], \quad (5)$$

$$\mathcal{L}_{pos}^D = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_d(\mathbf{x}, \mathbf{y})} [\log D_{y^p}^p(\mathbf{x})], \quad (6)$$

where D_i^d and D_i^p are the i^{th} element in D^d and D^p . For clarity, we will eliminate all subscripts for expected value notations, as all random variables are sampled from their respected distributions ($\mathbf{x}, \mathbf{y} \sim p_d(\mathbf{x}, \mathbf{y})$, $\mathbf{z} \sim p_z(\mathbf{z})$, $\mathbf{c} \sim p_c(\mathbf{c})$). The final objective for training D is the weighted average of all objectives:

$$\max_D \mathcal{L}^D = \lambda_g \mathcal{L}_{gan}^D + \lambda_d \mathcal{L}_{id}^D + \lambda_p \mathcal{L}_{pos}^D, \quad (7)$$

where we set $\lambda_g = \lambda_d = \lambda_p = 1$.

Meanwhile, G consists of an encoder G_{enc} and a decoder G_{dec} . G_{enc} aims to learn an identity representation $f(\mathbf{x}) = G_{enc}(\mathbf{x})$ from a face image \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} , and $\mathbf{z} \in \mathbb{R}^{N^z}$ is the noise modeling other variations 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 objectives:

$$\mathcal{L}_{gan}^G = \mathbb{E}[\log D^r(G(\mathbf{x}, \mathbf{c}, \mathbf{z}))], \quad (8)$$

$$\mathcal{L}_{id}^G = \mathbb{E}[\log D_{y^d}^d(G(\mathbf{x}, \mathbf{c}, \mathbf{z}))], \quad (9)$$

$$\mathcal{L}_{pos}^G = \mathbb{E}[\log D_{y^t}^p(G(\mathbf{x}, \mathbf{c}, \mathbf{z}))]. \quad (10)$$

Similarly, the final objective for training the discriminator G is the weighted average of each objective:

$$\max_G \mathcal{L}^G = \mu_g \mathcal{L}_{gan}^G + \mu_d \mathcal{L}_{id}^G + \mu_p \mathcal{L}_{pos}^G, \quad (11)$$

where we set $\mu_g = \mu_d = \mu_p = 1$.

G and D improves each other during the alternative training process. With D being more powerful in distinguishing real vs. fake images and classifying poses, G strives for synthesizing an identity-preserved face with the target pose to compete with D . We benefit from this process in three aspects. 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 as 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. Therefore, $f(\mathbf{x})$ 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 [55] with batch normalization (BN) for G_{enc} and D . Besides, since the stability of the GAN game suffers if sparse gradient layers (MaxPool, ReLU) are used, we replace them with strided convolution and exponential linear unit (ELU) respectively. D is trained to optimize Eqn. 7 by adding a fully connected layer with the softmax loss for real vs. fake, identity, and pose classifications respectively. G includes G_{enc} and G_{dec} that are bridged by the to-be-learned identity representation $f(\mathbf{x}) \in \mathbb{R}^{N^f}$, which is the AvgPool output in our G_{enc} . $f(\mathbf{x})$ is concatenated with a pose code \mathbf{c} and a random noise \mathbf{z} . A series of fractionally-strided convolutions (FConv) [20] transforms the $(N^f + N^p + N^z)$ -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. 11 when a synthetic face $\hat{\mathbf{x}}$ is fed to D and the gradient is back-propagated to update G .

Previous work in face rotation use $L2$ loss [13], [14] to enforce the synthetic face to be similar to the ground truth face at the

TABLE 1: The structures of G_{enc} , G_{dec} and D networks in single-image and multi-image DR-GAN. Blue texts represent extra elements to learn the coefficient ω in the G_{enc} of multi-image DR-GAN.

G_{enc} and D			G_{dec}		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
Conv11	$3 \times 3/1$	$96 \times 96 \times 32$	FC		$6 \times 6 \times 320$
Conv12	$3 \times 3/1$	$96 \times 96 \times 64$	FConv52	$3 \times 3/1$	$6 \times 6 \times 160$
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	$3 \times 3/1$	$6 \times 6 \times (N^f + 1)$	FConv11	$3 \times 3/1$	$96 \times 96 \times 3$
AvgPool	$6 \times 6/1$	$1 \times 1 \times (N^f + 1)$			
FC (D only)		$N^d + N^p + 1$			

target pose. This line of work requires the training data to include face image pairs of the same identity at different poses, which is achievable for controlled datasets such as Multi-PIE, but hard to fulfill for in-the-wild datasets. On contrary, DR-GAN does not require image pairs since there is no direct supervision on the synthetic images. This enables us to utilize extensive real-world unstructured datasets for model training. To initialize the training, given a training image, we randomly sample the pose code with equal probability for each pose view. Such a random sampling is conducted at *each* epoch during the training, for the purpose of assigning *multiple* pose codes to one training image. For the noise vector, we also randomly sample each dimension independently from the uniform distribution in the range of $[-1, 1]$.

3.3 Multi-Image DR-GAN

Our single-image DR-GAN extracts an identity representation and performs face rotation by processing one single image. 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.

The multi-image DR-GAN has the same D as single-image DR-GAN, but a different G as shown in Fig. 3. Given n images $\{\mathbf{x}_i\}_{i=1}^n$ of the same identity y^d at various poses as input, besides extracting the feature representation $f(\mathbf{x}_i)$, G_{enc} also estimates a confident coefficient ω_i for each image, which predicts the quality of the learnt representation. The fused representation of n images is the weighted average of all representations,

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{\sum_{i=1}^n \omega_i f(\mathbf{x}_i)}{\sum_{i=1}^n \omega_i}. \quad (12)$$

This fused representation is then 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 and a target pose y^t

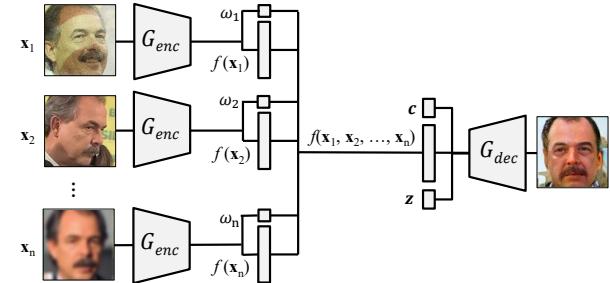


Fig. 3: Generator in multi-image DR-GAN. From an image set of a subject, we can fuse the features to a single representation via dynamically learnt coefficients and synthesize images in any pose.

specified by the pose code. Thus, each sub-objective for learning G has $(n + 1)$ terms:

$$\begin{aligned} \mathcal{L}_{gan}^G = \sum_{i=1}^n & [\mathbb{E}[\log(D^r(G(\mathbf{x}_i, \mathbf{c}, \mathbf{z})))]] \\ & + \mathbb{E}[\log(D^r(G(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{c}, \mathbf{z})))]. \end{aligned} \quad (13)$$

The similar extension applied for \mathcal{L}_{id}^G and \mathcal{L}_{pos}^G . The coefficient ω_i in Eqn. 12 is learned so that an image with a higher quality contributes more to the fused representation. The quality is an indicator of the PIFR performance of the image, 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 [47], [54]. Our coefficient learning is essentially the quality prediction, from novel perspectives in contrast to prior work. That is, without explicit supervision, it is driven by D through the decoded image $G_{dec}(f(\mathbf{x}_1, \dots, \mathbf{x}_n), \mathbf{c}, \mathbf{z})$, 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 and 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 convolutional filter 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 : each ω is set to 0 with a probability of 0.5. Hence, during training, the network takes any number of inputs varying from 1 to n .

DR-GAN can be used in PIFR, image quality prediction, and face rotation. While the network in Fig. 2 (d) is used for training, our network for testing is much simplified. First, for PIFR, only G_{enc} is used to extract the representation from one or multiple images. Second, for quality prediction, only G_{enc} is used to compute ω from one image. Thirdly, both G_{enc} and G_{dec} are used for face rotation by specifying a target pose and a noise vector.

3.4 Improving G_{enc} via Model Switch

The ultimate goal of DR-GAN is to learn a disentangled representation for PIFR. Our G_{enc} aims for identity representation learning. While our D^d aims for identity classification, it also learns an identity representation that could be used for face

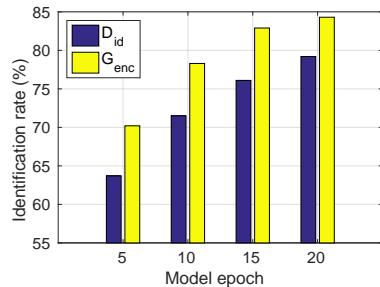


Fig. 4: Recognition performance of G_{enc} and D^d when training DR-GAN with different D^d on Multi-PIE dataset.

recognition during testing, the same as most previous work [55], [56]. The fact that both G_{enc} and D^d can be used for face recognition motivates us to explore two questions. First, whether G_{enc} can outperform D^d for PIFR. Second, whether a better D^d will lead to a better G_{enc} in representation learning.

To answer the above questions, we conduct a bounding experiment to compare the face recognition performance of G_{enc} and D^d . Specifically, using the Multi-PIE training set, we train a single-task CNN-based face recognition model for 20 epochs. We save the models at 5th, 10th, 15th, and 20th epochs, termed as D_5^d , D_{10}^d , D_{15}^d , D_{20}^d respectively. These four models can be used as D^d and to train four single-image DR-GAN models. Each model is trained until converged where we only update G with D^d being fixed, which leads to four G_{enc} termed as G_{enc}^5 , G_{enc}^{10} , G_{enc}^{15} , G_{enc}^{20} respectively.

Both G_{enc} and D^d are used to extract identity features for face recognition on Multi-PIE, with the results in Fig. 4. We have three observations. First, the performance of D^d shows that $D_5^d < D_{10}^d < D_{15}^d < D_{20}^d$. This is expected since the performance increases as the model is being trained for more epochs. Second, the performance of G_{enc} also shows a similar trend with $G_{enc}^5 < G_{enc}^{10} < G_{enc}^{15} < G_{enc}^{20}$, which indicates that a better D^d indeed leads to a better G_{enc} . Third, G_{enc} consistently outperforms D^d , which suggests that the learnt representation in G_{enc} is more discriminative than the representation in conventional CNN-based face recognition models.

Based on the above observations, we propose an iterative scheme to switch between G_{enc} and D^d in order to further improve G_{enc} . As shown in Tab. 1, G_{enc} and D^d shares the same network structure except that G_{enc} has an additional convolutional filter for the coefficient estimation. During training, we iteratively replace D^d with the latest G_{enc} by removing the additional convolutional filter after several epochs. Since G_{enc} can always outperform D^d , we will expect a better D^d after model switch. Moreover, a better D^d will lead to a better G_{enc} , which is then used as D^d for the next switch. This iterative switch will lead to a better representation and thus better PIFR performance.

3.5 Improving G_{dec} via Representation Interpolation

Our G_{enc} learns a mapping from the image space to a representation space and G_{dec} learns the mapping from the representation space to the image space. G_{enc} is important for PIFR while G_{dec} is crucial for face synthesis. The usage of pose code, random noise, as well as the model switch techniques are useful for learning a better disentangled representation for G_{enc} . However, even with a perfect representation from G_{enc} , a poor G_{dec} may synthesize unsatisfactory face images.

To learn a better G_{dec} , we propose to employ representation interpolation to regularize the learning process. Prior GANs [20] have observed that interpolation between two noise vectors can still produce a valid image. Similarly in our work, by assuming a convex identity space, the interpolation between two representations $f(\mathbf{x}_1)$, $f(\mathbf{x}_2)$ extracted from the face images \mathbf{x}_1 , \mathbf{x}_2 of two different identities should still be a valid face but with an unknown identity. During training, we randomly pair images with different identities to generate an interpolated representation:

$$f_\alpha(\mathbf{x}_1, \mathbf{x}_2) = \alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2). \quad (14)$$

We use the average, $f_{\frac{1}{2}}$, for simplicity. Other fixed or random weights can be used as well. Similar to the objectives for G and D in multi-image DR-GAN, we have additional terms to regularize the averaged representation. D aims to classify the generated image to the fake class by having the following extra term:

$$\mathbb{E}[\log(1 - D^r(G_{dec}(f_{\frac{1}{2}}(\mathbf{x}_1, \mathbf{x}_2), \mathbf{c}, \mathbf{z})))]. \quad (15)$$

And G aims to generate an image that can fool D to classify it as the real class and the target pose, and ignore the identity part, with two additional terms in \mathcal{L}_{gan}^G and \mathcal{L}_{pos}^G :

$$\mathbb{E}[\log(D^r(G_{dec}(f_{\frac{1}{2}}(\mathbf{x}_1, \mathbf{x}_2), \mathbf{c}, \mathbf{z})))], \quad (16)$$

$$\mathbb{E}[\log(D_{y^t}^p(G_{dec}(f_{\frac{1}{2}}(\mathbf{x}_1, \mathbf{x}_2), \mathbf{c}, \mathbf{z})))]. \quad (17)$$

With the proposed techniques to improve both G_{enc} and G_{dec} , we expect to improve the generalization ability of DR-GAN for both representation learning and image synthesis. As will be shown in the experiments, the proposed techniques are effective in improving the performance of DR-GAN.

3.6 Comparison to Prior GANs

We compare DR-GAN with most relevant GAN variants (Fig. 2). **Conditional GAN.** Conditional GAN [18], [57] extends GAN by feeding the labels to both G and D to generate images conditioned on labels, either class labels, 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.

Auxiliary Classifier GAN. Odena *et al.* [58] extends conditional GAN to add an additional classifier to D to classify real images into N^c classes. DR-GAN shares a similar loss for D but with a distinguish purpose. The auxiliary classifier in Odena *et al.* [58] is used to help improving the stability and quality of GAN training. Meanwhile, we employ two additional classifiers to guide the representation learning in the encoder-decoder structure G .

Adversarial Autoencoder (AAE). In AAE [59], 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 [59] 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.



Fig. 5: The mean faces of 13 pose groups in CASIA-Webface. The blurriness shows the challenges of pose estimation for large poses.

4 EXPERIMENTS

DR-GAN can be used for face recognition by using the learnt representation from G_{enc} , and face rotation by specifying different pose codes and noise vectors with G . We evaluate DR-GAN quantitatively for PIFR and qualitatively for face rotation. We further conduct experiments to analyze the training strategy, disentangle representation, and image coefficients. Our experiments are conducted for both controlled and in-the-wild databases.

4.1 Experimental Settings

Databases. Multi-PIE [28] is the largest database for evaluating face recognition under pose, illumination, and expression variations in controlled setting. For fair comparison, we follow the setting in [13]: using 337 subjects with neutral expression, 9 poses within $\pm 60^\circ$, and 20 illuminations. The first 200 subjects are used for training and the rest 137 subjects for testing. In the testing set, one image per subject with frontal view and neutral illumination forms the gallery set and the others are the probe set. For Multi-PIE experiments, we add an additional illumination code similar to the pose code to disentangle the illumination variation. Therefore, we have $N^d = 200$, $N^p = 9$, $N^{il} = 20$. Further, to demonstrate our ability in synthesizing large-pose faces, we train a second model with training faces up to 90° (i.e., $N^p = 13$).

For the in-the-wild setting, we train on CASIA-WebFace [55] and AFLW [60], and test on CFP [9] and IJB-A [23]. CASIA-WebFace includes 494, 414 near-frontal faces of 10, 575 subjects. We add the AFLW (25, 993 images) to the training set to supply more pose variation. Since there is no identity information in this dataset, those images only used to compute GAN, pose related losses. 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 $N^d = 10, 548$, $N^p = 13$. We set $N^f = 320$, $N^z = 50$ for both settings.

Implementation Details. Following [55], we align all face images to a canonical view of size 110×110 . We randomly sample 96×96 regions from the aligned 110×110 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 [61], [62] to classify each face to one of 13 poses. The mean face image of each pose group is shown in Fig. 5. The mean faces of profile faces are less sharp than those of the near-frontal pose groups, which indicates the pose estimation error caused by the face alignment algorithm.

Our implementation is extensively modified from a publicly available implementation of DC-GAN. We follow the optimization strategy in [20]. The batch size is set to be 64. All weights are initialized from a zero-centered normal distribution with a

TABLE 2: DR-GAN and its partial variants performance comparison.

Method	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN - D^r	80.0 ± 2.2	55.5 ± 3.5	88.7 ± 0.8	95.0 ± 0.8
DR-GAN - D^p	78.0 ± 2.0	53.9 ± 6.8	87.5 ± 0.8	94.5 ± 0.7
DR-GAN	81.2 ± 2.7	56.2 ± 9.1	89.0 ± 1.4	95.1 ± 0.9

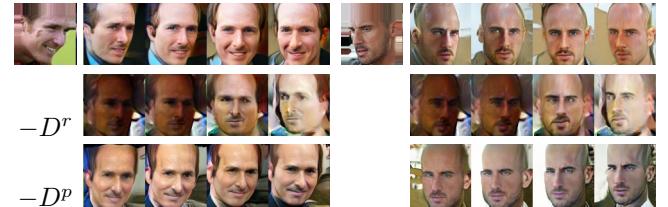


Fig. 6: Generated faces of DR-GAN and its partial variants.

standard deviation of 0.02. Adam optimizer [63] is used with a learning rate of 0.0002 and momentum 0.5.

Evaluation. The proposed DR-GAN aims for both face representation learning and face image synthesis. The cosine distance between two representations is used for face recognition. We also evaluate the performance of face recognition w.r.t. different numbers of images in both training and testing. For image synthesis, we show qualitative results by comparing different losses and interpolation of the learnt representations. We also evaluate the various effects of different components in our method.

4.2 Ablation study

Discriminator Components. Our discriminator is designed as a multi-task CNN with three components, namely D^g , D^d , D^p , for real/fake, identity and pose classification respectively. While D^d plays a critical role to guide the generator to preserve the input identity, we would like to study the role of the remaining components. Table 2 presents the recognition performance of single-image DR-GAN partial variants with each of D components removed. While the variant without adversarial loss has a slightly performance drop, the model without pose classification task has more severe drop. This shows the important of generating face images in different poses. Also, the role of each component is shown in generated faces (Fig. 6). When removing D^r , generated images has lower quality although they can be realized as faces and in correct poses. When removing D^p , the pose of generated images can't be controlled by the pose code and usually affected by the input face's pose. This can be caused by pose information residing in the feature representation. This also explains the severe drop in the model's recognition performance.

Disentangled Representation. In DR-GAN, we claim that the learnt representation is disentangled from pose variations via the pose code. To validate this, following the energy-based weight visualization method proposed in [56], we perform feature visualization on the FC layer, denoted as $\mathbf{h} \in \mathbb{R}^{6 \times 6 \times 320}$, in G_{dec} . Our goal is to select two out of the 320 filters that have highest

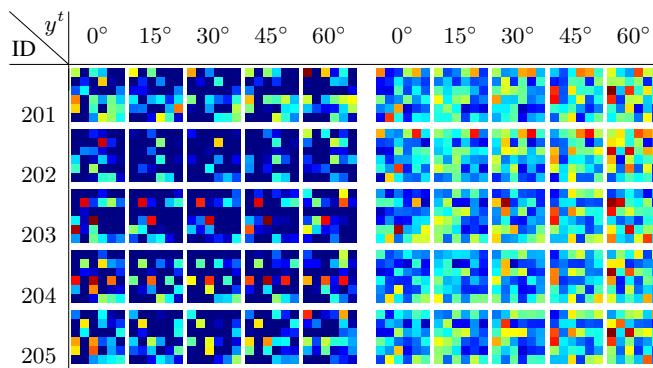


Fig. 7: Responses of two filters: filter with the highest responses to identity (left), and pose (right). Responses of each row are of the same subject, and each column are of the same pose. Note the within-row similarity on the left and within-column similarity on the right.

responses for identity and pose respectively. The assumption is that if the learnt representation is pose-invariant, there should be separate neurons to encode the identity features and pose features.

Recall that we concatenate $f(\mathbf{x}) \in \mathbb{R}^{320}$, $\mathbf{c} \in \mathbb{R}^{13}$ and $\mathbf{z} \in \mathbb{R}^{50}$ into one feature vector, which multiplies with a weight matrix $\mathbf{W}_{fc} \in \mathbb{R}^{(320+13+50) \times (6 \times 6 \times 320)}$ and generates the output \mathbf{h} with $\mathbf{h}^i \in \mathbb{R}^{6 \times 6}$ being the feature output of one filter in FC. Let $\mathbf{W}_{fc} = [\mathbf{W}_{fx}; \mathbf{W}_c; \mathbf{W}_z]$ denote the weight matrix with three sub-matrices, which would multiply with $f(\mathbf{x}), \mathbf{c}, \mathbf{z}$ respectively. Taking the identity matrix as an example, we have $\mathbf{W}_{fx} = [\mathbf{W}_{fx}^1, \mathbf{W}_{fx}^2, \dots, \mathbf{W}_{fx}^{320}]$ where $\mathbf{W}_{fx}^i \in \mathbb{R}^{320 \times 36}$. We compute an energy vector $\mathbf{s}_d \in \mathbb{R}^{320}$ with each element as: $s_d^i = \|\mathbf{W}_{fx}^i\|_F$. We then find the filter with the highest energy in \mathbf{s}_d as $k_d = \arg \max_i s_d^i$. Similarly, by partitioning \mathbf{W}_c , we find another filter, denoted as k_p , with the highest energy for pose.

Given the representation $f(\mathbf{x})$ of one subject, along with a pose code \mathbf{c} and noise \mathbf{z} , we can compute the responses of two filters via $\mathbf{h}^{k_d} = (f(\mathbf{x}); \mathbf{c}; \mathbf{z})^\top \mathbf{W}_{fc}^{k_d}$ and $\mathbf{h}^{k_p} = (f(\mathbf{x}); \mathbf{c}; \mathbf{z})^\top \mathbf{W}_{fc}^{k_p}$. By varying the subjects and pose codes, we generate two arrays of responses in Fig. 7, for identity (\mathbf{h}^{k_d}) and pose (\mathbf{h}^{k_p}) respectively. For both arrays, each row represents the responses of the same subject and each column represents the same pose. The responses for identity encode the identity features, where each row shows similar patterns and each column does not share similarity. On contrary, for pose responses, each column share similar patterns while each row is not related. This visualization supports our claim that the learnt representation is pose-invariant.

Single vs. Multiple Image DR-GAN. We evaluate the effect of the number of training images (n) per subject on the face recognition performance on CFP. 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 face verification performance on CFP using $f(\mathbf{x})$ of each model are shown in Tab. 3. We observe the advantage of multi-image DR-GAN over the single-image counterpart despite they use the *same amount* of training data, 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.

Model Switch. In Sec. 3.4, we propose to improve G_{enc} via model switch, i.e., replacing D^d with G_{enc} during training. Table 4 shows the performance of G_{enc} for face recognition on Multi-PIE. At the beginning, G_{enc} is initialized with a model

TABLE 3: Comparison of single vs. multi-image DR-GAN on CFP.

Method	Frontal-Frontal	Frontal-Profile
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	93.41 ± 1.17

TABLE 4: Performance of G_{enc} on Multi-PIE when keep switching to D_d . At Epoch 0, G_{enc} is trained with only the softmax loss.

Epoch No.	0	20	40	60	80	100
Identification rate (%)	79.7	84.8	87.1	88.7	89.8	90.4

trained with the softmax loss for identity classification. We use G_{enc} to replace D^d and retrain G with random initialization. When G converges, based on the accuracy on a 5-held-out-subjects validation set, we replace D^d with G_{enc} and repeat above steps. Table 4 reports face recognition performance of G_{enc} on Multi-PIE test set at each switch. Clearly, the performance keeps improving as training goes on. This study implies that DR-GAN may leverage the future development of face recognition, by using a 3rd party recognizer as D^d and further improve upon it.

4.3 Confident Coefficients

In multi-image DR-GAN, we learn a confident coefficient for each input image by assuming that the learnt coefficient is indicative of the image quality, i.e., how good it can be used for face recognition. Therefore, a low-quality image should have a relatively poor representation and small coefficients so that it would contribute less to the fused representation. To validate this assumption, we compute the confident coefficients for all images in IJB-A and CFP databases and plot the distribution as shown in Fig. 8.

For IJB-A, we show four example images with low, medium-low, medium-high, and high coefficients. It is obvious that the learnt coefficients are correlated to the image quality. Images with relatively low coefficients are usually blurring, with large poses or failure cropping. While images with relatively high coefficients are of very high quality with frontal faces and less occlusion. Since CFP consists of 5,000 frontal faces and 2,000 profile faces, we plot their distributions separately. Despite some overlap in the middle region, the profile faces clearly have relatively low coefficients compared to the frontal faces. Within each distribution, the coefficient are related to other variations expect yaw angles. The low-quality images for each pose group are with occlusion and/or challenging lighting conditions, while the high-quality ones are with less occlusion and under normal lighting.

To quantitatively evaluate the correlation between the coefficients and face recognition performance, we conduct an identity classification experiment on IJB-A. Specifically, we randomly select all frames of one video for each subject and select half of images for training and remaining for testing. The training and testing sets share the same identities. Therefore, in the testing stage, we can use the output of the softmax layer as the probability of each testing image belonging to the right identity class. This probability is an indicator of how well the input image can be recognized as the true identity. Given the estimated coefficients, we plot these two values for the testing set, as shown in Fig. 9. These two values are highly correlated to each other with a correlation of 0.69, which again supports our assumption that the learnt coefficients are indicative of the image quality.

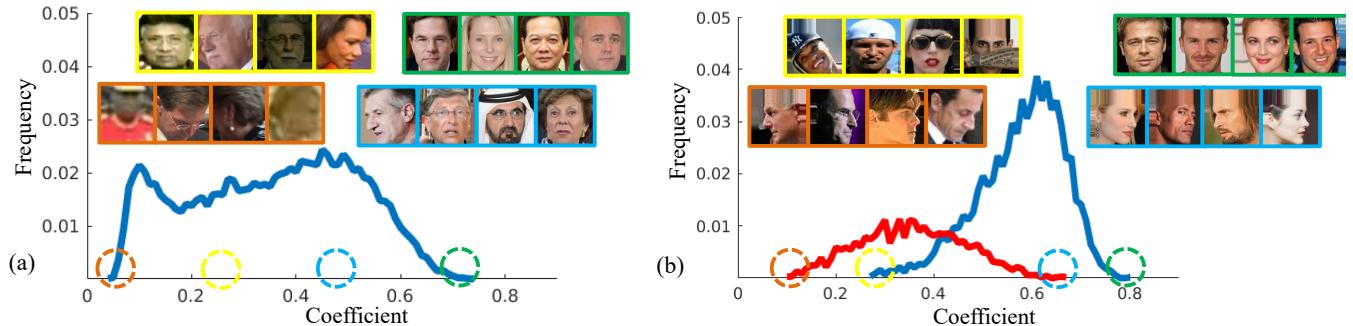


Fig. 8: Coefficient distributions on IJB-A (a) and CFP (b). For IJB-A, we visualize images at four regions of the distribution. For CFP, we plot the distributions for frontal faces (blue) and profile faces (red) separately and show images at the heads and tails of each distribution.

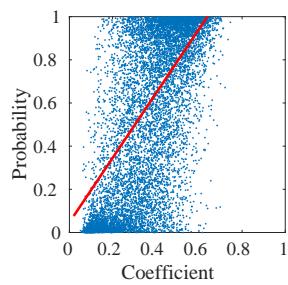


Fig. 9: Coefficients and classification probabilities correlation.

TABLE 5: Performance of IJB-A when removing images by threshold ω_t . “Selected” shows the percentage of retained images.

ω_t	Selected (%)	Verification		Identification	
		@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
0	100.0	84.3 ± 1.4	72.6 ± 4.4	91.0 ± 1.5	95.6 ± 1.1
0.1	94.9	84.2 ± 1.7	72.7 ± 2.9	91.3 ± 1.3	95.7 ± 1.0
0.25	71.9	83.6 ± 1.2	73.3 ± 3.0	90.7 ± 1.2	95.2 ± 1.0
0.5	24.6	80.9 ± 1.9	71.3 ± 4.7	86.5 ± 1.9	93.1 ± 1.6
1.0	5.7	77.8 ± 2.2	64.0 ± 6.2	83.4 ± 2.3	91.6 ± 1.2

Image selection with ω . One common application of image quality is to prevent low-quality images from contributing to face recognition. To validate whether our coefficients have such usability, we design the following experiment. For each template in IJB-A, we keep images whose coefficients ω are larger than a predefined threshold ω_t , or if all ω are smaller we keep one image with the highest ω . Tab. 5 reports the performance on IJB-A, with different ω_t . With ω_t being 0, all test images are kept and the result is the same as Tab. 8. These results show that keeping all or majority of the samples are better than removing them. This is encouraging as it reflects the effectiveness of DR-GAN in automatically diminishing the impact of low-quality images, without removing them by thresholding.

Feature fusion with ω . We also would like to show our proposed feature fusion using coefficient ω is effective for the template to template matching purpose. We compare it with multiple fusion methods in both feature level and score level. Table 6 shows comparisons of different fusion methods on our multi-image DR-GAN features. To compare two template with size n_1, n_2 , for score-level, min, max, mean are respectively taking minimum, maximum and average of all n_1n_2 possible pairwise distances. Mean-min is the average of $n_1 + n_2$ minimum distances from

TABLE 6: Fusion schemes comparisons on IJB-A dataset.

Method	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
Score	Min	78.3 ± 2.7	46.0 ± 6.9	86.7 ± 1.4
	Max	22.8 ± 2.0	12.3 ± 2.3	30.6 ± 2.8
	Mean	72.8 ± 2.9	49.2 ± 5.3	85.7 ± 1.3
	Mean-min	82.4 ± 2.2	58.5 ± 6.3	90.2 ± 1.0
	Softmax	84.3 ± 1.6	69.2 ± 6.8	90.1 ± 1.0
Feature	Max	19.0 ± 1.3	12.1 ± 1.7	45.4 ± 5.3
	Mean	83.0 ± 1.5	67.0 ± 4.8	89.6 ± 1.5
	ω -fusion	84.3 ± 1.4	72.6 ± 4.4	91.0 ± 1.5
ω -fusion				95.6 ± 1.1

each feature from one template to the other. All of these methods have the time complexity of $\mathcal{O}(n_1n_2)$. Softmax, proposed in [64], aggregates multiple weighted averages of the pair-wise scores, where each weight is the function of the score using an exponential function in different scales. It has the time complexity of $\mathcal{O}(mn_1n_2)$, where m is the number of weight scale. Here, following [15], we use a total of $m = 21$ scales from 0 to 20. For feature-level fusion, max, mean are respectively max-pooling and average-pooling along each feature dimension. All feature-level fusion methods, including our ω -fusion, have the time complexity of $\mathcal{O}(n_1+n_2)$. From Tab. 6, our fusion using estimated ω achieves the best performance among all methods.

4.4 Representation Learning

Loss Function Comparison. Our G_{dec} and D can be viewed as a loss function for $f(\mathbf{x})$. Typical loss functions used in deep learning-based face recognition can be divided into two categories: probability- and energy-based losses. Probability-based losses (i.e., softmax and its variants) usually compute a distribution of probability to all identities. Meanwhile, energy-based losses (contrastive, triplet, etc.) associate an energy to each configuration. Here, we compare DR-GAN to multiple common loss functions of face recognition. To have a fair comparison on IJB-A, for all functions, we use our G_{enc} network architecture and “mean min” fusion. DR-GAN by itself can surpass all prior loss functions (Tab. 7). Also, any advanced loss function can also be beneficial to DR-GAN: energy-based losses (center, triplet, etc.) can be employed directly on our representation $f(\mathbf{x})$ or probability-based losses (angular, additive-margin softmax, etc.) can be used to replace the D_d ’s softmax. Empirically, using additive-margin softmax [65] as a softmax replacement on D_d can further improve DR-GAN performance, we name this variant as DR-GAN_{AM}.

TABLE 7: Loss function comparisons. All use “mean min” fusion.

Method	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
Softmax	75.9 \pm 3.9	44.1 \pm 9.9	87.8 \pm 0.9	94.6 \pm 0.6
Center [66]	74.9 \pm 3.1	50.3 \pm 7.0	87.2 \pm 1.4	95.2 \pm 0.9
Triplet [3]	74.9 \pm 3.1	50.3 \pm 7.0	87.2 \pm 1.4	95.2 \pm 0.9
AM-Softmax [65]	81.3 \pm 3.0	52.7 \pm 8.9	88.7 \pm 0.7	94.3 \pm 0.4
DR-GAN _{single img.}	81.2 \pm 2.7	56.2 \pm 9.1	89.0 \pm 1.4	95.1 \pm 0.9
DR-GAN	82.4 \pm 2.3	58.5 \pm 8.0	90.2 \pm 1.0	95.6 \pm 0.5
DR-GAN _{AM}	85.7 \pm 1.6	70.3 \pm 5.79	91.0 \pm 1.5	95.6 \pm 1.1

TABLE 8: Performance comparison on IJB-A dataset.

Method	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
GOTS [23]	40.6 \pm 1.4	19.8 \pm 0.8	44.3 \pm 2.1	59.5 \pm 2.0
Wang <i>et al.</i> [25]	72.9 \pm 3.5	51.0 \pm 6.1	82.2 \pm 2.3	93.1 \pm 1.4
DCNN [24]	78.7 \pm 4.3	—	85.2 \pm 1.8	93.7 \pm 1.0
PAM _{frontal} [15]	73.3 \pm 1.8	55.2 \pm 3.2	77.1 \pm 1.6	88.7 \pm 0.9
PAMs [15]	82.6 \pm 1.8	65.2 \pm 3.7	84.0 \pm 1.2	92.5 \pm 0.8
p-CNN [56]	77.5 \pm 2.5	53.9 \pm 4.2	85.8 \pm 1.4	93.8 \pm 0.9
FF-GAN [67]	85.2 \pm 1.0	66.3 \pm 3.3	90.2 \pm 0.6	95.4 \pm 0.5
DR-GAN [27]	77.4 \pm 2.7	53.9 \pm 4.3	85.5 \pm 1.5	94.7 \pm 1.1
DR-GAN _{split}	84.3 \pm 1.4	72.6 \pm 4.4	91.0 \pm 1.5	95.6 \pm 1.1
DR-GAN _{split+inter}	85.6 \pm 1.5	75.1 \pm 4.2	91.3 \pm 1.6	95.8 \pm 1.0
DR-GAN _{AM}	87.2 \pm 1.4	78.1 \pm 3.5	92.0 \pm 1.3	96.1 \pm 0.7

TABLE 9: Performance (Accuracy) comparison on CFP.

Method	Frontal-Frontal	Frontal-Profile
Sengupta <i>et al.</i> [9]	96.40 \pm 0.69	84.91 \pm 1.82
Sankarana <i>et al.</i> [68]	96.93 \pm 0.61	89.17 \pm 2.35
Chen <i>et al.</i> [69]	98.67 \pm 0.36	91.97 \pm 1.70
Human	96.24 \pm 0.67	94.57 \pm 1.10
DR-GAN [27]	97.84 \pm 0.79	93.41 \pm 1.17
DR-GAN _{split+inter}	98.13 \pm 0.81	93.64 \pm 1.51
DR-GAN _{AM}	98.36 \pm 0.75	93.89 \pm 1.39

Results on Benchmark Databases. We compare DR-GAN with state-of-the-art face recognizers on IJB-A, CFP and Multi-PIE.

Table 8 shows the performance of both face identification and verification on IJB-A. For our results, we report results of multi-image DR-GAN using the proposed ω -fusion. The first row shows the performance of our preliminary work [27]. “split” represents the model trained with the separated D^r . “+inter” represents the additional changes made by the representation interpolation proposed in Sec. 3.5, which is shown to be effective in improving the face recognition performance. The final row presents the variant using additive margin softmax [65] (also with “split” and “interpolation”). Compared to the state of the art, DR-GAN achieves superior results on both verification and identification. Also, our work has made substantial improvement over the preliminary version [27]. These in-the-wild results show the power of DR-GAN for PIFR.

Table 9 shows the comparison on CFP evaluated with Accuracy. Results are reported with the average with standard deviation over 10 folds. Overall, we achieve comparable performance on frontal-frontal verification while having 1.92% improvement on the frontal-profile verification.

Table 10 shows the face identification performance on Multi-

TABLE 10: Identification rate (%) comparison on Multi-PIE dataset.

Method	0°	15°	30°	45°	60°	Average
Zhu <i>et al.</i> [70]	94.3	90.7	80.7	64.1	45.9	72.9
Zhu <i>et al.</i> [13]	95.7	92.8	83.7	72.9	60.1	79.3
Yim <i>et al.</i> [14]	99.5	95.0	88.5	79.9	61.9	83.3
Using L_2 loss	95.1	90.8	82.7	72.7	57.9	78.3
DR-GAN [27]	97.0	94.0	90.1	86.2	83.2	89.2
DR-GAN	98.1	94.9	91.1	87.2	84.6	90.4
DR-GAN _{AM}	98.1	95.0	91.3	88.0	85.8	90.8

TABLE 11: Representation $f(\mathbf{x})$ vs. synthetic image $\hat{\mathbf{x}}$ on IJB-A.

Features	Verification		Identification	
	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
$f(\hat{\mathbf{x}})$	78.5 \pm 1.9	60.3 \pm 3.7	86.9 \pm 1.6	94.2 \pm 1.3
$D^d(\hat{\mathbf{x}})$	77.1 \pm 2.9	53.5 \pm 6.2	85.7 \pm 1.7	93.6 \pm 1.6
$f'(\hat{\mathbf{x}})$	79.2 \pm 2.9	60.8 \pm 7.3	89.2 \pm 1.4	95.3 \pm 1.1
$f'(\hat{\mathbf{x}}) \& f(\hat{\mathbf{x}})$	83.0 \pm 1.8	71.7 \pm 3.6	90.7 \pm 1.4	95.6 \pm 1.0
$f(\mathbf{x})$	84.3 \pm 1.4	72.6 \pm 4.4	91.0 \pm 1.5	95.6 \pm 1.1

PIE compared to the methods with the same setting. Our method shows a significant improvement for large-pose faces, e.g., there is more than 20% improvement margin at $\pm 60^\circ$ poses. The variation of recognition rates across different poses is much smaller than the baselines, which suggests that our learnt representation is more robust to the pose variation.

Representation vs. Synthetic Image for PIFR. Many prior work [10], [11] use frontalized faces for PIFR. To evaluate the identity preservation of synthetic images from DR-GAN, we also perform face recognition using our frontalized faces. Any face feature extractor could be applied to them, including G_{enc} or D^d . However, both are trained on real images of various poses. To specialize to synthetic frontal faces, we fine-tune G_{enc} with the synthetic images and denote as $f'(\cdot)$. As shown in Tab. 11, although the performance of synthetic images (and its score-level fusion denoted as $f'(\hat{\mathbf{x}}) \& f(\hat{\mathbf{x}})$) is not as good as the learnt representation, using the fine-tuned G_{enc} on synthetic frontal still achieves comparable performance to the previous methods, which shows the identity preservation ability of DR-GAN.

4.5 Face Rotation

Adversarial Loss vs. L_2 loss. Prior work [14], [35], [70] on face rotation normally employ the L_2 loss to learn a mapping between two views. To compare the L_2 loss with our adversarial loss, we train a model where G is supervised by an L_2 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. 10, DR-GAN can generate far more realistic faces that are similar to the ground truth faces in all views. Meanwhile, images synthesized by the L_2 loss cannot maintain high frequency components and are blurry. In fact, L_2 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. 10. In contrast, by integrating the adversarial loss, we expect to learn a more discriminative representation for better recognition, and a more generative representation for better face synthesis.

Variable Interpolations. 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.



Fig. 10: Face rotation comparison on Multi-PIE. Given the input (in illumination 07 and 75° pose), we show synthetic images of L_2 loss (top), adversarial loss (middle), and ground truth (bottom). Column 2-5 show the ability of DR-GAN in simultaneous face rotation and re-lighting.

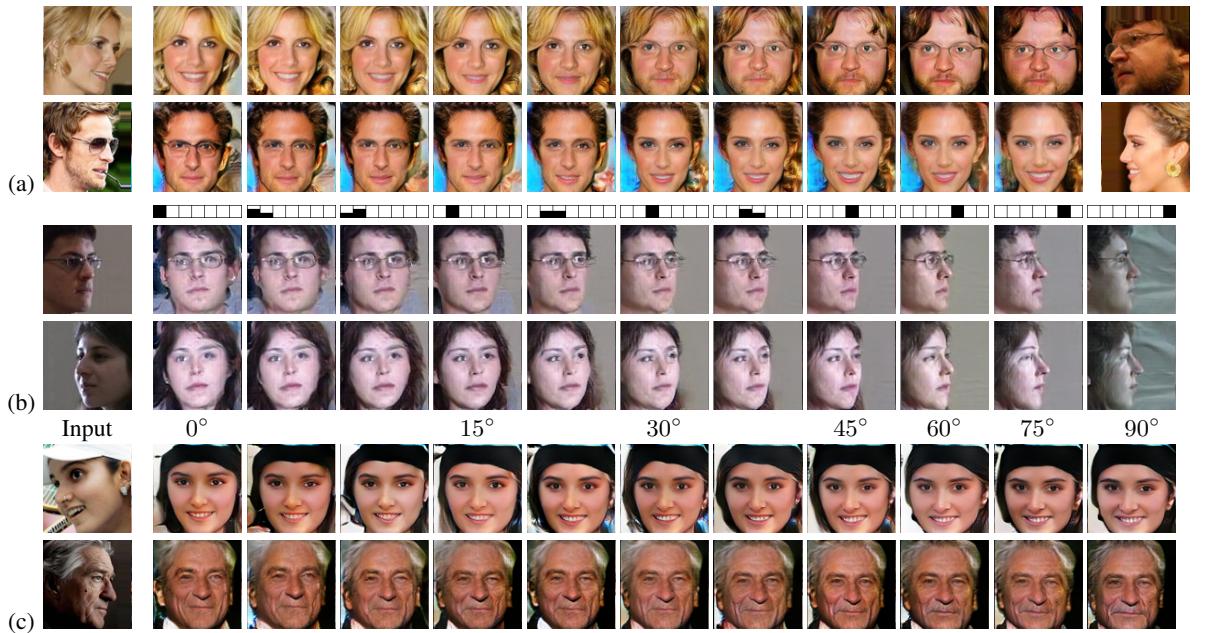


Fig. 11: Interpolation of $f(\mathbf{x})$, \mathbf{c} , and \mathbf{z} . (a) Interpolating between identity representations of two faces (Column 1 and 12). Note the smooth transition between different genders and facial attributes. (b) Interpolating in-between *unseen* poses via *continuous* pose codes, shown above Row 3 ($0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$ are available in training). (c) Interpolating between $\mathbf{z} = -1$ (Column 2) and $\mathbf{z} = 1$ (Column 12).

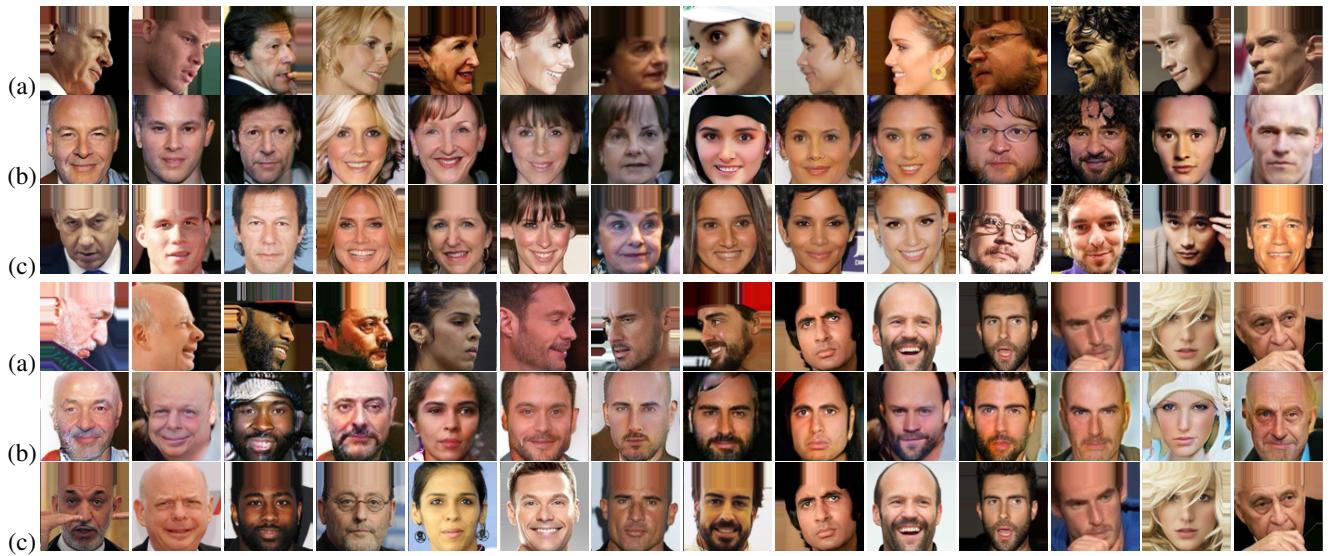


Fig. 12: Face frontalization on CFP: (a) input, (b) frontalized faces, (c) real frontal faces. We expect the frontalized faces to preserve the identity, rather than all facial attributes. This is very challenging for face rotation due to the in-the-wild variations and extreme profile views. Also, synthetic faces can remove some variations (roll, expression, or occlusions) for frontal input.

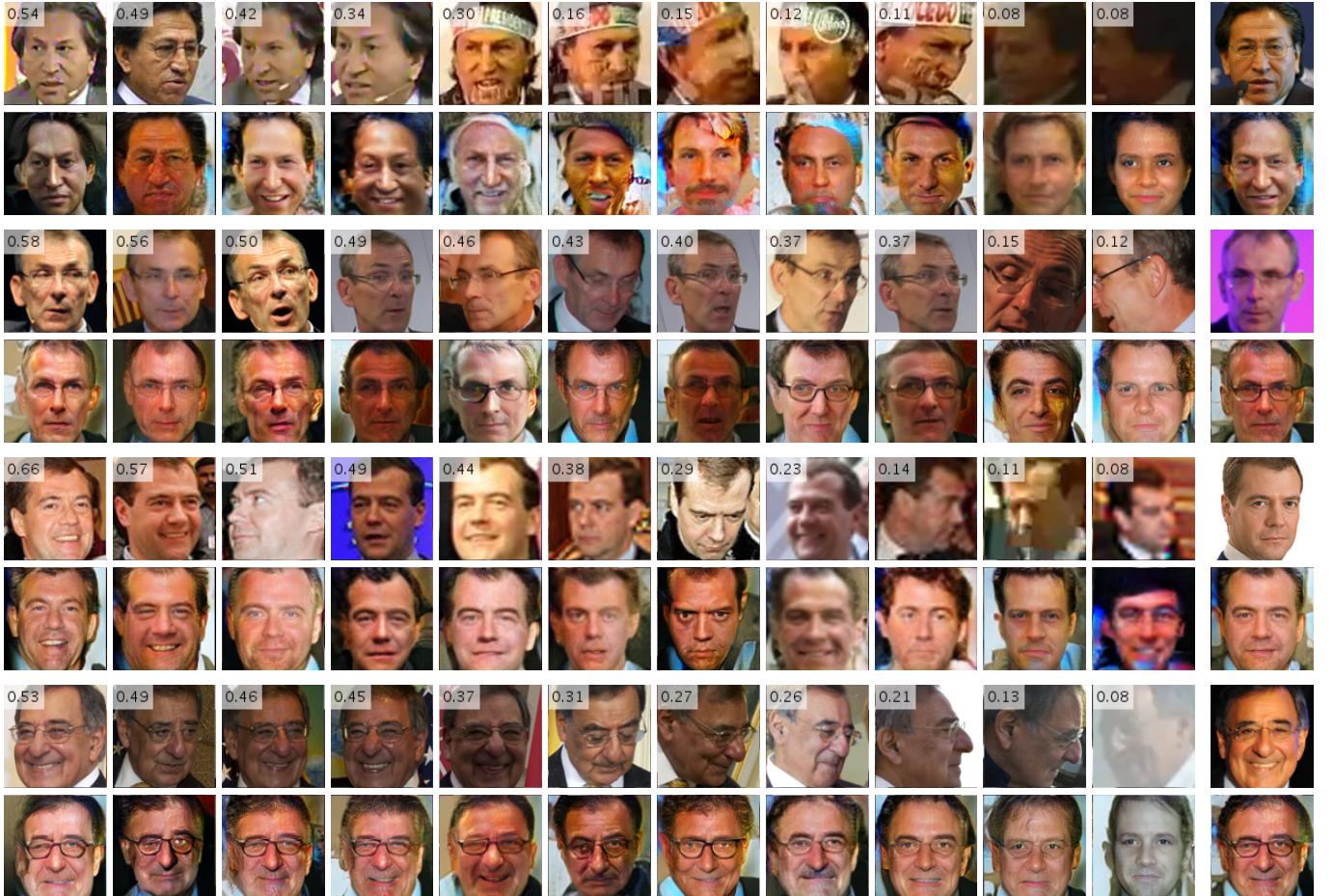


Fig. 13: Face frontalization on IJB-A. For each of four subjects, we show 11 input images with estimated coefficients overlaid at the top left corner (first row) and their frontalized counter part (second row). The last column is the groundtruth frontal and synthetic frontal from the fused representation of all 11 images. Note the challenges of large poses, occlusion, and low resolution, and our *opportunistic* frontalization.



Fig. 14: Face frontalization on IJB-A for an image set (first subject) and a video sequence (second subject). For each subject, we show 11 input images (first row), their respective frontalized faces (second row) and the frontalized faces using *incrementally* fused representations from all previous inputs up to this image (third row). In the last column, we show the groundtruth frontal face.

In Fig. 11 (a), the top row shows a transition from a female subject to a male subject with beard and glasses. Similar to [20], these smooth semantic changes indicate that the model has learned essential identity representations for image synthesis.

Similar interpolation can be conducted for the pose codes as well. During training, we use a one-hot vector \mathbf{c} to specify the *discrete* pose of the synthetic image. During testing, we could generate face images with *continuous* poses, whose pose code is the weighted average, i.e., interpolation, of two neighboring pose codes. Note that the resultant pose code is no longer a one-hot vector. As in Fig. 11 (b), this leads to smooth pose transition from one view to many views *unseen* to the training set.

We can also interpolate the noise vector \mathbf{z} . We synthesize frontal faces at $\mathbf{z} = -1$ and $\mathbf{z} = 1$ (a vector of all 1s) and interpolate between two \mathbf{z} . Given the fixed identity representation and pose code, the synthetic images are identity-preserved frontal faces. As in Fig. 11 (c), the change of \mathbf{z} leads to the change of the background, illumination condition, and facial attributes such as beard, while the identity is well preserved and faces are of the frontal view. Thus, \mathbf{z} models less significant face variations.

Face Rotation on Benchmark Databases. Our generator is trained to be a face rotator. Given one or multiple face images with arbitrary poses, we can generate multiple identity-preserved faces at different views. Figure 10 shows the face rotation results on Multi-PIE. Given an input image at any pose, we can generate multi-view images of the same subject but at a different pose by specifying different pose codes or in a different lighting condition by varying illumination code. The rotated faces are similar to the ground truth with well-preserved attributes such as eyeglasses.

One application of face rotation is face frontalization. Our DR-GAN can be used for face frontalization by specifying the frontal-view as the target pose. Figure 12 shows the face frontalization on CFP. Given an extreme profile input image, DR-GAN can generate a realistic frontal face that has similar identity characteristics as the real frontal face. To the best of our knowledge, this is the first work that is able to *frontalize a profile-view in-the-wild face image*. When the input image is already in the frontal view, the synthetic images can correct the pitch and roll angles, normalize illumination and expression, and impute occluded facial areas, as shown in the last few examples of Fig. 12.

Figure 13 shows face frontalization results on IJB-A. For each subject or template, we show 11 images and their respective frontalized faces, and the frontalized face generated from the fused representation. For each input image, the estimated coefficient ω is shown on the top-left corner of each image, which clearly indicates the quality of the input image as well as the frontalized image. For example, coefficients for low-quality or large-pose input images are very small. These images will have very little contribution to the fused representation. Finally, the face from the fused representation has superior quality compared to all frontalized images from a single input face. This shows the effectiveness of our multi-image DR-GAN in taking advantage of multiple images of the same subject for better representation learning.

To further evaluate face frontalization results w.r.t. different numbers of input images, we vary the number of input images from 1 to 11 and visualize the frontalized images from the *incrementally* fused representations. As shown in Fig. 14, the individually frontalized faces have varying degrees of resemblance to the true subject, according to the qualities of different input images. The synthetic images from fused representations (third row) improve as the number of images increases.

5 CONCLUSIONS

This paper presents DR-GAN to learn a disentangled representation for PIFR, by modeling the face rotation process. We are the first to construct the generator in GAN with an encoder-decoder structure for representation learning, which can be quantitatively evaluated by performing PIFR. Using the pose code for decoding and pose classification in the discriminator lead to the disentanglement of pose variation from the identity features. We also propose multi-image DR-GAN to leverage multiple images per subject in both training and testing to learn a better representation. This is the first work that is able to frontalize an extreme-pose in-the-wild face. We attribute the superior PIFR and face synthesis capabilities to the discriminative yet generative representation learned in G . Our representation is discriminative since the other variations are explicitly disentangled by the pose/illumination codes, and random noise, and is generative since its decoded (synthetic) image would still be classified as the original identity.

REFERENCES

- [1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *CVPR*, 2014.
- [2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *BMVC*, 2015.
- [3] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015.
- [4] X. Liu and T. Chen, “Pose-robust face recognition using geometry assisted probabilistic modeling,” in *CVPR*, 2005.
- [5] X. Liu, J. Rittscher, and T. Chen, “Optimal pose for face recognition,” in *CVPR*, 2006.
- [6] X. Chai, S. Shan, X. Chen, and W. Gao, “Locally linear regression for pose-invariant face recognition,” *TIP*, 2007.
- [7] R. Abiantun, U. Prabhu, and M. Savvides, “Sparse feature extraction for pose-tolerant face recognition,” *TPAMI*, 2014.
- [8] C. Ding and D. Tao, “A comprehensive survey on pose-invariant face recognition,” *TIST*, 2016.
- [9] 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.
- [10] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in *CVPR*, 2015.
- [11] 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.
- [12] M. Kan, S. Shan, H. Chang, and X. Chen, “Stacked Progressive Auto-Encoders (SPAЕ) for face recognition across poses,” in *CVPR*, 2014.
- [13] 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.
- [14] 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.
- [15] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, “Pose-aware face recognition in the wild,” in *CVPR*, 2016.
- [16] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *TMM*, 2015.
- [17] 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.
- [18] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv:1411.1784*, 2014.
- [19] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, “Deep generative image models using a Laplacian pyramid of adversarial networks,” in *NIPS*, 2015.
- [20] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *ICLR*, 2016.
- [21] 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.
- [22] D. Berthelot, T. Schumm, and L. Metz, “BEGAN: Boundary Equilibrium Generative Adversarial Networks,” *arXiv:1703.10717*, 2017.
- [23] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A,” in *CVPR*, 2015.

- [24] J.-C. Chen, V. M. Patel, and R. Chellappa, “Unconstrained face verification using deep CNN features,” in *WACV*, 2016.
- [25] D. Wang, C. Otto, and A. K. Jain, “Face search at scale,” *TPAMI*, 2016.
- [26] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, “Do we really need to collect millions of faces for effective face recognition?” in *ECCV*, 2016.
- [27] L. Tran, X. Yin, and X. Liu, “Disentangled Representation Learning GAN for pose-invariant face recognition,” in *CVPR*, 2017.
- [28] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-PIE,” *IVC*, 2010.
- [29] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *ICML*, 2016.
- [30] X. Yu and F. Porikli, “Ultra-resolving face images by discriminative generative networks,” in *ECCV*, 2016.
- [31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in *NIPS*, 2016.
- [32] A. Odena, “Semi-supervised learning with generative adversarial networks,” in *ICMLW*, 2016.
- [33] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, “Morphable displacement field based image matching for face recognition across pose,” in *ECCV*, 2012.
- [34] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, “Robust statistical face frontalization,” in *ICCV*, 2015.
- [35] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, “Weakly-supervised disentangling with recurrent transformations for 3D view synthesis,” in *NIPS*, 2015.
- [36] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu, “Random faces guided sparse many-to-one encoder for pose-invariant face recognition,” in *ICCV*, 2013.
- [37] J. Roth, Y. Tong, and X. Liu, “Unconstrained 3D face reconstruction,” in *CVPR*, 2015.
- [38] ——, “Adaptive 3D face reconstruction from unconstrained photo collections,” *TPAMI*, 2017.
- [39] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *TPAMI*, 2013.
- [40] R. Marc’Aurelio, F. J. Huang, Y.-L. Boureau, and Y. LeCun, “Unsupervised learning of invariant feature hierarchies with applications to object recognition,” in *CVPR*, 2007.
- [41] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, “Deep convolutional inverse graphics network,” in *NIPS*, 2015.
- [42] S. Bharadwaj, M. Vatsa, and R. Singh, “Biometric quality: A review of fingerprint, iris, and face,” *EURASIP JIVP*, 2014.
- [43] P. Grother and E. Tabassi, “Performance of biometric quality measures,” *TPAMI*, 2007.
- [44] Y. Tong, F. W. Wheeler, and X. Liu, “Improving biometric identification through quality-based face and fingerprint biometric fusion,” in *CVPRW*, 2010.
- [45] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, “Design and evaluation of photometric image quality measures for effective face recognition,” *IET Biometrics*, 2014.
- [46] M. Abdel-Mottaleb and M. H. Mahoor, “Application notes-algorithms for assessing the quality of facial images,” *IEEE Computational Intelligence Magazine*, 2007.
- [47] N. Ozay, Y. Tong, F. W. Wheeler, and X. Liu, “Improving face recognition with a quality-based probabilistic framework,” in *CVPRW*, 2009.
- [48] Y. Chen, S. C. Dass, and A. K. Jain, “Localized iris image quality using 2-D wavelets,” in *ICB*, 2006.
- [49] E. Krichen, S. Garcia-Salicetti, and B. Dorizzi, “A new probabilistic iris quality measure for comprehensive noise detection,” in *BTAS*, 2007.
- [50] E. Tabassi and C. L. Wilson, “A novel approach to fingerprint image quality,” in *ICIP*, 2005.
- [51] R. Teixeira and N. Leite, “A new framework for quality assessment of high-resolution fingerprint images,” *TPAMI*, 2016.
- [52] D. Muramatsu, Y. Makihara, and Y. Yagi, “View transformation model incorporating quality measures for cross-view gait recognition,” *IEEE transactions on cybernetics*, 2016.
- [53] D. S. Matovski, M. Nixon, S. Mahmoodi, and T. Mansfield, “On including quality in applied automatic gait recognition,” in *ICPR*, 2012.
- [54] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition,” in *CVPRW*, 2011.
- [55] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv:1411.7923*, 2014.
- [56] X. Yin and X. Liu, “Multi-task convolutional neural network for face recognition,” August 2017.
- [57] H. Kwak and B.-T. Zhang, “Ways of conditioning generative adversarial networks,” in *NIPSW*, 2016.
- [58] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” in *ICML*, 2017.
- [59] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, “Adversarial autoencoders,” in *ICLRW*, 2015.
- [60] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.
- [61] A. Jourabloo and X. Liu, “Pose-Invariant 3D face alignment,” in *ICCV*, 2015.
- [62] ——, “Large-Pose face alignment via CNN-based dense 3D model fitting,” in *CVPR*, 2016.
- [63] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
- [64] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, R. Nevatia, and G. Medioni, “Face recognition using deep multi-pose representations,” in *WACV*, 2016.
- [65] F. Wang, W. Liu, H. Liu, and J. Cheng, “Additive margin softmax for face verification,” *arXiv preprint arXiv:1801.05599*, 2018.
- [66] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European Conference on Computer Vision*, 2016.
- [67] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, “Towards large-pose face frontalization in the wild,” in *In Proceeding of International Conference on Computer Vision*, Venice, Italy, October 2017.
- [68] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, “Triplet probabilistic embedding for face verification and clustering,” in *BTAS*, 2016.
- [69] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, “Fisher vector encoded deep convolutional features for unconstrained face verification,” in *ICIP*, 2016.
- [70] Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep learning identity-preserving face space,” in *ICCV*, 2013.

Luan Tran received his B.S. in Computer Science from Michigan State University with High Honors in 2015. He is now pursuing his Ph.D. also at Michigan State University in the area of deep learning and computer vision. His areas of interest include applying generative models for representation learning. He is a member of the IEEE.



Xi Yin received the B.S. degree in Electronic and Information Science from Wuhan University, China, in 2013. Since August 2013, she has been working toward her Ph.D. degree in the Department of Computer Science and Engineering, Michigan State University, USA. Her research area are face recognition, deep learning, and plant image processing. Her paper on multi-leaf segmentation won the Best Student Paper Award at Winter Conference on Application of Computer Vision (WACV) 2014.



Xiaoming Liu is an Assistant Professor at the Department of Computer Science and Engineering of Michigan State University. He received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric (GE) Global Research. His research interests include computer vision, machine learning, and biometrics. As a co-author, he is a recipient of Best Industry Related Paper Award runner-up at ICPR 2014, Best Student Paper Award at WACV 2012 and 2014, and Best Poster Award at BMVC 2015. He has been the Area Chair for numerous conferences, including FG, ICPR, WACV, ICIP, and CVPR. He is the program co-chair of WACV 2018 and BTAS 2018. He is an Associate Editor of Neurocomputing journal. He has authored more than 100 scientific publications, and has filed 26 U.S. patents.