IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2020, accepted July 8, 2020, date of publication July 15, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009512

Dual Encoder-Decoder Based Generative Adversarial Networks for Disentangled Facial Representation Learning

CONG HU[1,2,3], ZHENHUA FENG[4,5], (Member, IEEE), XIAOJUN WU[1,2], (Member, IEEE), AND JOSEF KITTLER[5], (Life Member, IEEE)

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2]Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China
[3]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China
[4]Department of Computer Science, University of Surrey, Guildford GU2 7XH, U.K.
[5]Center for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K.

Corresponding author: Xiaojun Wu (wu_xiaojun@jiangnan.edu.cn)

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) Programme under Grant (FACER2VM) EP/N007743/1, in part by the National Natural Science Foundation of China under Grant U1836218, Grant 61672265, Grant 61876072, and Grant 61902153, in part by the 111 Project of Chinese Ministry of Education under Grant B12018, and in part by the Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, under Grant MJUKF-IPIC202002.

ABSTRACT To learn disentangled representations of facial images, we present a Dual Encoder-Decoder based Generative Adversarial Network (DED-GAN). In the proposed method, both the generator and discriminator are designed with deep encoder-decoder architectures as their backbones. To be more specific, the encoder-decoder structured generator is used to learn a pose disentangled face representation, and the encoder-decoder structured discriminator is tasked to perform real/fake classification, face reconstruction, determining identity and estimating face pose. We further improve the proposed network architecture by minimizing the additional pixel-wise loss defined by the Wasserstein distance at the output of the discriminator so that the adversarial framework can be better trained. Additionally, we consider face pose variation to be continuous, rather than discrete in existing literature, to inject richer pose information into our model. The pose estimation task is formulated as a regression problem, which helps to disentangle identity information from pose variations. The proposed network is evaluated on the tasks of pose-invariant face recognition (PIFR) and face synthesis across poses. An extensive quantitative and qualitative evaluation carried out on several controlled and in-the-wild benchmarking datasets demonstrates the superiority of the proposed DED-GAN method over the state-of-the-art approaches.

INDEX TERMS Disentangled representation learning, encoder-decoder, generative adversarial networks, face synthesis, pose invariant face recognition.

I. INTRODUCTION

Benefiting from the rapid development of deep learning and the easy access to a large number of annotated face images, face recognition [1]–[4] has advanced significantly in recent years. Although impressive performance has been achieved on several benchmarking databases, pose variation is still one of the crucial bottlenecks for many practical applications [5], [6]. Facial appearance variations caused by poses are even larger than those caused by different identities [7]. To mitigate this difficulty, many approaches have been proposed for pose-invariant face recognition (PIFR). Existing PIFR methods can be divided into three categories. One approach is to remap non-frontal faces to frontal ones, and then extract facial features from frontalised faces for better face representation [8]–[12]. The second one is to learn pose-invariant representations directly from non-frontal faces [13]–[16]. The last category aims to learn disentangled facial representations so that identity-preserving features can be disentangled from pose variation [17], [18]. Our proposed method belongs to the last category.

The consensus regarding desirable properties of good representations of data has recently been established in [19]–[22]. Disentanglement, one of the properties of good

The associate editor coordinating the review of this manuscript and approving it for publication was Peter Peer.

representation, is a kind of distributed feature representation in which disjoint dimensions of a latent code reflect different high-level generative factors of data. The disentanglement is also often described as statistical independence; each independent factor is expected to be semantically well aligned with the human intuition regarding the data generative factors. Specifically, the disentangled representation can separate explanatory factors that interact non-linearly in real-world data, such as object shapes, material properties, light sources and so on. A representation distilling each important factor of data into a single independent direction is hard to learn, but it is highly valuable for many other downstream tasks like PIFR and face synthesis across views [23]–[26].

Deep generative models facilitate learning disentangled representations. It is a methodology that enables learning of the probability distribution of data and generating new samples according to control codes in a latent space. By learning the appropriate parameters, deep generative models can generate new data mimicking the distribution of the target data. Once a disentangled representation is learned, the disjoint dimensions of the hidden code model the data generative factors separately. These underlying factors have the potential to explain the major variations in the data. When only one factor varies but all others are fixed, the generated sequence of samples can show an interpretable change to human beings. For example, when we generate a hand-written digit, a component of the code may be associated with the stroke width. When its value is changed, only the stroke width of the generated digit becomes smaller, while other factors on the images (e.g. class, shape, color) stay the same. In recent years, Variational Auto-Encoder (VAE) [27] and Generative Adversarial Networks (GAN) [28] based methods as two notable branches of deep generative model have successfully been used in the disentangled representation learning. For instance, β-VAE [29] learns disentangled latent codes by encouraging the latent distribution to be close to the standard normal distribution, in which each random variable is independent. DC-IGN [30] is another VAE-based generative model for disentangled representation learning. However, DC-IGN may not apply to unstructured in-the-wild images, since it achieves disentanglement by providing batch training samples with one attribute being fixed. InfoGAN [31] also uses statistical independence, which is motivated by the principle of maximization of mutual information. The Disentangled Representation learning GAN (DR-GAN) [18] learns generative and discriminative facial representations, which disentangle the face identity from pose so that it can better handle cross-pose recognition. DR-GAN is also similar to the prior work [10] in which joint representation learning and face rotation are explored with a multi-task CNN. In summary, most of the existing works disentangle the factors by using statistical independence of a prior distribution.

Although DR-GAN has achieved impressive performance in face synthesis across poses and PIFR, it has some problems: 1) The process of training of DR-GAN is not stable. In a few stable cases, a mode collapse often occurs, producing degenerate images; 2) The pose variations are categorized into several distinct classes by a one-hot vector. Consequently, although it is a strong prior, the pose information is insufficient for disentangled facial representation learning. To improve the training stability of GAN, the encoder-decoder structured discriminator has been successfully used in EBGAN [32] and BEGAN [33], which is also used as a backbone network in our method. To achieve stable model training, an equilibrium enforcing method was proposed in BEGAN, in which a hyper-parameter is introduced to balance the generator and discriminator during the model training. Different from the classical GANs, BEGAN aims to match the auto-encoder loss distributions, not between sample distributions. We also introduce an equilibrium enforcing strategy in our method. However, in contrast to BEGAN, our method not only matches the distributions between samples like in typical GANs, but also the distributions of the reconstruction losses of samples, which is conducive to better representation learning. Accordingly, pixel-wise reconstruction error is used as another loss function, aside identity loss and pose estimation in our GAN model.

DR-GAN codes the pose into several classes with a one-hot vector, incurring information loss in the process. Pose changes continuously, non-linearly but smoothly. For this reason, we represent pose code by a continuous variable rather than in a discrete form. This also allows estimating the pose by regression rather than classification.

This paper addresses the problem of learning a generative model for disentangled facial representation extraction. By combining the advanced techniques of GAN-based representation learning methods, we propose to learn disentangled pose-robust features by modeling the complex non-linear transform between face images with different poses through a dual encoder-decoder structured deep neural network in an adversarial way, namely Dual Encoder-Decoder based Generative Adversarial Networks (DED-GAN). The proposed network is evaluated in terms of the quality of face synthesis of different views on the one hand and pose-invariant face recognition (PIFR) on the other hand. Our contributions are summarised as follows:

- A new GAN architecture with fast and stable convergence is proposed for disentangled facial representation learning.
- Our proposed method can generate a face with arbitrary pose variations.
- The proposed method learns identity-preserving features simultaneously.
- To the best of our knowledge, this is the first attempt to use pose regression for disentangled face representation. The proposed continuous pose variation model provides more detailed information about the pose. It is used explicitly to control the manifold of identity-preserving face synthesis.

- Experiments in PIFR and face synthesis across poses demonstrate the advantage of our method on multiple benchmarking databases.

The rest of the paper is organised as follows: We first overview the existing literature related to the proposed method in Section II. Then we present the proposed DED-GAN in Section III and introduce the implementation details in Section IV. An ablation study and experimental results are reported in Section V. Last, the conclusion is drawn in Section VI.

II. RELATED WORK

A. GENERATIVE ADVERSARIAL NETWORK

Recently, the state-of-the-art in deep generative models, especially in VAE [27] and GAN [28], have advanced significantly. As one of the most promising deep neural networks, GAN has attracted widespread attention from the computer vision and machine learning communities. It provides a simple, yet powerful way to estimate data distribution and generate realistic samples by the zero-sum two-player game [34]. Through modeling a real sample distribution, a GAN can encourage the generated samples to move towards the true image manifold, and thus generate photo-realistic images with plausible high-frequency details. However, the classical GAN suffers from computational problems, *e.g.* the inferior performance caused by unbalanced training of the generator without comparable attention given to updating the discriminator. A collapsed generator will lose the capacity to fit the target data distribution. To address the aforementioned model collapse issue, some improved GAN architectures have been proposed. For example, Zhao *et al.* [32] proposed energy-based GAN (EBGAN) that considers the generator and discriminator as energy functions. Salimans *et al.* [35] introduced a bag of tricks to address GAN training strategies and achieved great performance on semi-supervised learning. Karras *et al.* [36] used a strategy of progressively growing the generator and discriminator of a GAN for improved image generation quality, stability and variation. Further, Arjovsky *et al.* [37] presented Wasserstein GAN (WGAN) using the earth mover's distance. They proved that WGAN is able to avoid the mode collapse problem to a certain extent.

Existing GAN models can handle most of the challenging cases, in which the pose, illumination and expression of faces are unconstrained. For example, Radford *et al.* [38] designed DC-GAN that evaluates a set of constraints on the architectural topology of convolutional GANs, which make the model stable to train. Huang *et al.* [39] focused on the local patches that have some semantic meaning and proposed TP-GAN. Li *et al.* [40] focused on the missing parts of the face and came up with a novel two adversarial losses as well as a semantic parsing loss to complete the faces. He *et al.* [41] edited the face images with desired attributes while preserving other details by encoder-decoder structured GAN. Both [42] and [43] applied an extension of GAN to a conditional setting and showed their utility in many tasks, including image in-painting [44], super-resolution [45], style transfer [46], face attribute manipulation [47] and even data augmentation for classification models [48], [49]. The VariGAN model was proposed by Zhao *et al.* [50] to solve the problem of generating multi-view images from a single viewpoint. Tran *et al.* [51] put forward DR-GAN, which fuses the pose information and can take one or multiple face images with yaw angles as input to achieve pose invariant facial representation learning. Similarly, Antipov *et al.* [52] concentrated on improving face synthesis in cross-age scenarios. Considering scene structure and context, Yang *et al.* [53] presented LR-GAN that learns generated image background and foreground separately and recursively to produce a completely natural or face image.

These successful GANs provide a strong motivation to learn disentangled facial representation and to develop a method for different view synthesis. However, there are several crucial issues with GANs such as training being unstable and a quantitative evaluation proving difficult. The previous work either focuses on the stability of training, the task of synthesising images, or using the features in the discriminator for image recognition. In contrast, we propose an innovative method for constructing the generator for disentangled representation learning, which is stable. The proposed DED-GAN method is also quantitatively evaluated for pose invariant face recognition.

B. POSE INVARIANT REPRESENTATION LEARNING

In conventional face recognition methods, local descriptors [54]–[57] and metric learning [58], [59] are often used to tackle the effect of pose variation. In contrast, deep learning methods handle pose variation through building pose-specific or pose-agnostic models with specific loss functions [60], [61]. For instance, the DeepFace [62] model uses a deep CNN coupled with 3D face alignment. The inception architecture, utilised in FaceNet [15], is used in DeepID2+ [63] and DeepID3 [64] where multi-task learning and metric learning are performed simultaneously. However, such data-driven methods heavily rely on well-annotated data. Collecting labeled data covering all variations is time-consuming and labor-intensive. Our proposed Dual Encoder-Decoder based GAN (DED-GAN) presents an idea similar to Disentangled Representation learning GAN (DR-GAN) [18], which considers both face rotation and representation learning in a unified network. However, our proposed model differs from DR-GAN in the following aspects: 1) we use a continuous pose code for disentangling face representation in DED-GAN, as it provides more detailed information about the pose as a strong prior for training, and 2) DR-GAN suffers from poor generalisation and from optimisation difficulties, which limit its effectiveness in face synthesis and face recognition. In contrast, our DED-GAN overcomes these issues by disentangling the pose utilizing pose regression and adding face reconstruction as a side task.

III. THE PROPOSED APPROACH

Our Dual Encoder-Decoder based GAN (DED-GAN) model learns two tasks simultaneously: synthesis of

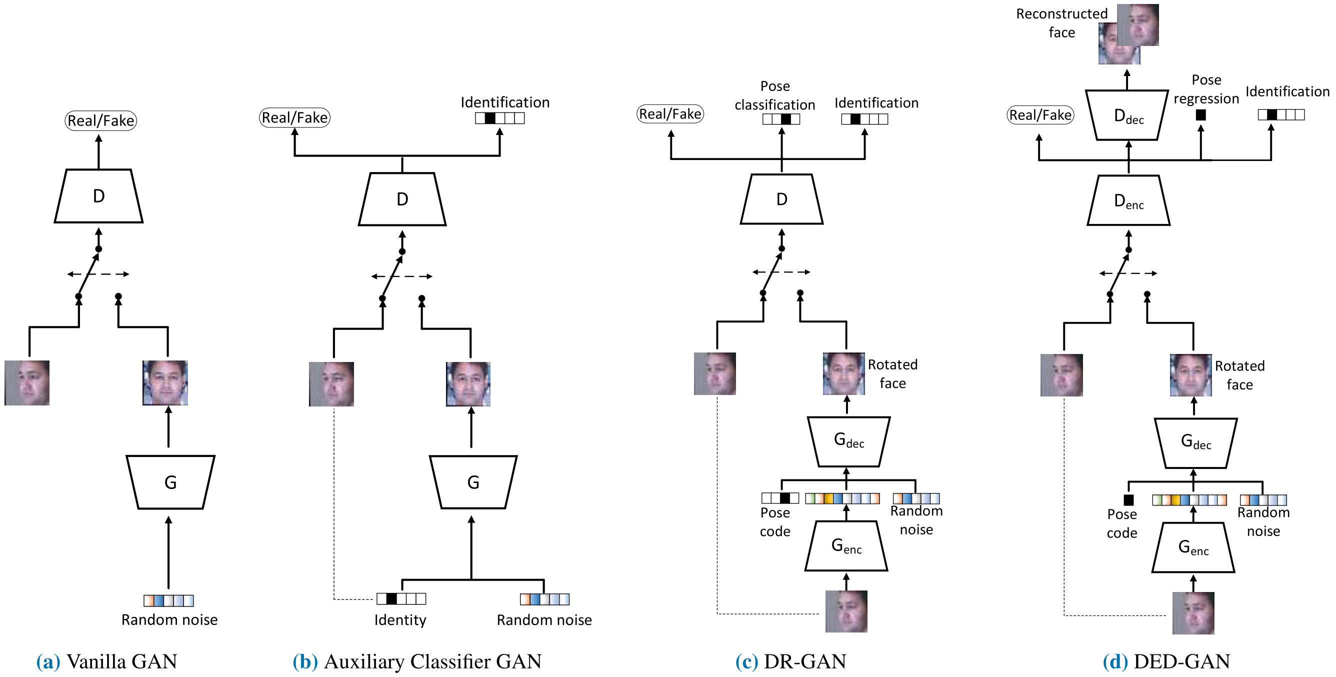

FIGURE 1. **Comparison of previous GANs architecture and our proposed DED-GAN.**

different face poses and pose-invariant face recognition. The encoder-decoder structured generator is used for face rotation and untangling the identity from pose variation. The encoder-decoder structured discriminator is used for facial reconstruction, pose estimation, identity classification and real/fake adversarial learning. The architecture of our DED-GAN is shown in Fig. 1d. We also show different architectures of earlier GANs such as Vanilla GAN, Auxiliary Classifier GAN and DR-GAN for comparison in Fig. 1a, Fig. 1b and Fig. 1c. In contrast to DR-GAN, we add a decoder to the discriminator, which is optimised for pixel-wise loss defined in terms of the Wasserstein distance, to balance the generator and discriminator. We also code the pose using a continuous variate instead of the discrete variate commonly specified by a one-hot vector. As a result, the task of pose disentanglement in the discriminator can be formulated as one of pose regression instead of classification, which further benefits the learning process.

It should be noted that the Encoder-Decoder structured discriminator has also been successfully used in BEGAN [33], to match the pixel-wise loss distributions of reconstructed real and synthesised samples. Our method also incorporates an Encoder-Decoder as the backbone of the discriminator to achieve a balanced learning behavior as part of weakly adversarial learning. Different from previous GANs, including BEGAN, our method not only sets out to match data distributions but also attempts to match image reconstruction loss distributions. This is achieved by using a typical GAN objective combined with an additional equilibrium term. To provide a detailed description of our approach, we start by introducing the original GAN, followed by our proposed DED-GAN method.

A. GENERATIVE ADVERSARIAL NETWORK

A typical GAN model consists of two networks pitted against one another in a two-player game: a generative model, G, is trained to synthesise images resembling the real data distribution and a discriminative model, D, is trained to distinguish the samples synthesised by G and real ones from the training data. The generator generates unlabelled realistic samples from the latent variable model to improve the discriminative ability of the discriminator. To learn the generator's distribution p_g over data x, we define a prior on input noise variables $p_z(z)$. The mapping $G(z;\theta_g)$ of z into the data space is achieved by a neural network with parameters θ_g, where G is a differentiable function. A second neural network with parameters θ_d is defined by $D(x;\theta_d)$ that outputs a single scalar. $D(x)$ represents the probability that x comes from the real data, p_d, rather than p_g. We train D to maximize the probability of assigning the correct label to both training examples and samples from G. We simultaneously train G to minimise $log(1-D(G(z)))$. In other words, the generator and discriminator are fighting against each other, which can be formulated as:

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

where z denotes a random noise, typically sampling from a Gaussian normal distribution, p_z. $G(z)$ denotes a sample synthesised by the generator and p_d denotes the distribution of real data. It is proved in the original GAN [28] that this minimax game has a global optimum when the distribution p_g of the synthetic samples converges to the distribution p_d of the real samples. At the beginning of training, the samples generated by G are extremely poor and thus they are rejected by D with high confidence. This minimax game theoretically has a global optimum for $p_g = p_d$. G and D are trained to alternatively optimise the following objectives:

$$\begin{aligned}\max_D L &= E_{x\sim p_d(x)}[logD(x)]\\ &\quad +E_{z\sim p_z(z)}[log(1-D(G(z)))], \qquad (2)\\ \min_G L &= E_{z\sim p_z(z)}[log(1-D(G(z)))]. \qquad (3)\end{aligned}$$

After several steps of the optimisation process, the generator and discriminator will reach the point at which neither can improve because $p_g = p_d$. The discriminator is unable to differentiate between the two distributions, *i.e.* $D(x) = 1/2$.

B. DUAL ENCODER-DECODER BASED GAN

Our DED-GAN explicitly disentangles face imaging factors to obtain an interpretable face representation for PIFR and face synthesis across poses. The backbone of DED-GAN consists of an encoder-decoder based generator and encoder-decoder based discriminator, as depicted in Fig. 1d. It learns the representation of a face by using the generator, where the encoded output of the generator is the identity-preserving representation. The representation is one part of the input to the decoder to synthesise various faces of the same subject with different attributes, *i.e.*, by virtually rotating the facial pose code. We not only match the distribution of face images by using classical real/fake adversarial learning, but also the distributions of the reconstruction error of samples reconstructed from the representation by using pixel-wise adversarial learning. As numerous variations manifest in face images such as pose, illumination and expression influence face recognition even more than changes in identity, it is desirable to prevent the generator from generating different facial representations for the same person with different face poses. In this work, we focus on pose variations and disentangle the pose information as an explicit variation. This facilitates learning a truly discriminative face representation.

1) PROBLEM FORMULATION

Our method aims to train a generative adversarial model conditioned on the real face image x and specified pose code c. Given a face image x with label $y = \{y^a, y^d, y^c\}$, where y^a, y^d and y^c represent the labels for real/fake, identity and pose. There are two tasks in our learning method: to learn a disentangled identity representation for PIFR and to synthesise faces across poses with different pose code c.

Different from the discriminator in the original GAN, our discriminator could be seen as a multi-task CNN consisting of four components: $D = [D^a, D^d, D^c, D^r]$, where $D^a \in \mathbb{R}^1$ is for classical real/fake adversarial learning, $D^d \in \mathbb{R}^{N^d}$ is for identity classification with N^d as the total number of subjects in the training set, $D^r \in \mathbb{R}^{N^c*N^w*N^h}$ is for face reconstruction and $D^c \in \mathbb{R}^{N^1}$ is for pose regression.

For the pose regression task, we first obtain the pose coefficients of all the training images. To obtain the pose of an image, we use the MTCNN method to extract 5 facial landmarks for each face image [65]. Then we transform face landmarks to the pose code using a statistical shape model [66]. Mathematically, we can express the face shape with a base shape s_0 plus a linear combination of n shape eigenvectors s_i as:

$$s = s_0 + \sum_{i=1}^{n} c_i s_i, \qquad (4)$$

where s_0 is the mean shape, s_i is the ith shape eigenvector by applying principal component analysis to all the training shapes and c_i is the corresponding coefficient. In general, the first shape eigenvector controls pose variations of the model thus we use c_1 as the pose code c.

The discriminator aims to classify the face image x as real or fake, to maximize the gap between the reconstruction error of real image and that of the synthetic image, and to estimate its identity and pose. Given an input image x, a random pose code c and a random noise z, the generator G generates a synthesised face image $G(x, c, z)$. The discriminator D attempts to classify the image using the following objectives:

$$\begin{aligned}L^D_{adv} &= E_{x,y\sim p_d(x,y)}[-logD^a(x)]\\ &\quad +E_{x,y\sim p_d(x,y),z\sim p_z(z),c\sim p_c(c)}\\ &\quad \times [-log(1-D^a(G(x,c,z)))], \qquad (5)\\ L^D_{id} &= E_{x,y\sim p_d(x,y)}[-logD^d_{y^d}(x)], \qquad (6)\\ L^D_{pos} &= E_{x,y\sim p_d(x,y)}|D^c_{y^c}(x)|, \qquad (7)\\ L^D_{pixel} &= E_{x\sim p_d(x),z\sim p_z(z),c\sim p_c(c)}\\ &\quad |D^r(x) - k\cdot D^r(G(x,c,z))|. \qquad (8)\end{aligned}$$

where k is a trade-off parameter to balance the distribution of reconstruction error of real faces and that of synthetic faces. For clarity, we eliminate all subscripts for expected value notation, as all random variables are sampled from their respected distributions $(x, y) \sim p_d(x, y), z \sim p_z(z), c \sim p_c(c)$. D^d is used for identity classification. It should be noted that pose regression D^c is used here rather than pose classification. The final objective for training D is the weighted average of all objectives:

$$minL^D = \lambda_a L^D_{adv} + \lambda_d L^D_{id} + \lambda_c L^D_{pos} + \lambda_r L^D_{pixel}, \qquad (9)$$

where λ_a, λ_d, λ_c and λ_r denote the weights of the four losses.

The generator G consists of an encoder G_{enc} and a decoder G_{dec}, where G_{enc} aims to learn an identity-preserving representation $f(x) = G_{enc}(x)$ from a face image x, G_{dec} is tasked to synthesise a face image $G_{dec}(f(x), c, z)$ with identity y^d and a target pose specified by c, and $z \in R^{N^z}$ is a noise variable, modelling other variations besides identity or pose. The pose code $c \in R^1$ is of continuous value. The goal of G is to fool D to classify $G(x, c, z)$ to the identity of input x and estimate the target pose with the following objectives:

$$\begin{aligned}L^G_{adv} &= E_{x,y\sim p_d(x,y),z\sim p_z(z),c\sim p_c(c)}[-logD^a(G(x,c,z))], \qquad (10)\\ L^G_{id} &= E_{x,y\sim p_d(x,y)}[-logD^d_{y^d}(G(x,c,z))], \qquad (11)\\ L^G_{pos} &= E_{x,y\sim p_d(x,y)}|D^c_{y^c}(G(x,c,z))|, \qquad (12)\\ L^G_{pixel} &= E_{x\sim p_d(x),z\sim p_z(z),c\sim p_c(c)}|D^r(G(x,c,z))|. \qquad (13)\end{aligned}$$

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

$$minL^G = \mu_a L^G_{adv} + \mu_d L^G_{id} + \mu_c L^G_{pos} + \mu_r L^G_{pixel}. \tag{14}$$

where μ_a, μ_d, μ_c and μ_r denote the weights of the four losses.

2) PIXEL-WISE LOSS

While classical GANs try to match data distributions directly with L_{adv}, our method additionally aims to match auto-encoder loss distributions using a pixel-wise loss L_{pixel} based on Wasserstein distance. Firstly, we introduce the auto-encoder loss, and then we compute a lower bound to the Wasserstein distance between the auto-encoder loss distributions of real and generated samples.

Let $L : R^{N_x} \mapsto R^+$, denote the loss for training a pixel-wise auto-encoder defined as:

$$L(x) = |x - D(x)|^\eta \tag{15}$$

where $D : R^{N_x} \mapsto R^{N_x}$ is the auto-encoder, $\eta \in \{1, 2\}$ is the target norm, and $x \in R^{N_x}$ is a sample of dimension N_x. Furthermore, let $\mu_{1,2}$ be two distributions of auto-encoder losses, and $\Gamma(\mu_1, \mu_2)$ be the set all of couplings of μ_1 and μ_2, whose respective means are $m_{1,2} \in R$. The Wasserstein distance can be expressed as:

$$W_1(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1,\mu_2)} E_{(x_1,x_2)\sim\gamma}[|x_1, x_2|] \tag{16}$$

Using Jensen's inequality, we can derive a lower bound to $W_1(\mu_1, \mu_2)$:

$$\inf E[|x_1, x_2|] \geq \inf |E[x_1 - x_2]| = |m_1 - m_2| \tag{17}$$

We design the discriminator to maximise $|m_1 - m_2|$ by forcing $m_1 \to 0, m_2 \to \infty$. Given the discriminator and generator parameters θ_D and θ_G, each to be updated by minimising the losses L^D_{pixel} and L^G_{pixel}, we express the optimisation problem in terms of a pixel-wise loss function:

$$L^D_{pixel} = L(x) - k_t \cdot L(G(x)) \tag{18}$$

$$L^G_{pixel} = L(G(x)) \tag{19}$$

$$k_{t+1} = k_t + \lambda_k(\beta L(x) - L(G(x))) \tag{20}$$

where k_t controls how much emphasis is put on $L(G(x))$ during gradient descent, λ_k is the learning rate for k. β is diversity ratio as a hyper-parameter to balance $L(x)$ and $L(G(x))$.

IV. IMPLEMENTATION DETAILS

The proposed Dual Encoder-Decoder based GAN (DED-GAN) is composed of a generator G and a discriminator D. Both are based on deep encoder-decoder networks. We follow the design for making G in the DR-GAN. The modified CASIA Net [67] is used as the backbone network. It consists of five convolution blocks, including one double-convolution block and four triple-convolution blocks, followed by an average pooling (AvePool) layer for feature extraction.

The generator G is composed of an encoder G_{enc} and a decoder G_{dec}, *i.e.*, $G = [G_{enc}, G_{dec}]$. Given a face image x, the encoder's output code $e = G_{enc}(x) \in R^{N_e}$ from the AvePool layer is concatenated with a pose code $c \in R^{N_c}$ and a noise $z \in R^{N_z}$ to form $[e, c, z]$, which is used as the input of G_{dec}. G_{dec} is a de-convolution neural network that transforms $[e, c, z]$ to a decoded face image, *i.e.*, $\hat{x} = G_{dec}([e, c, z])$. D_a and D_r are used to force the distributions of both synthesised samples and their auto-encoder losses to match those of real samples. The discriminator D is composed of an encoder D_{enc} and a decoder D_{dec}, *i.e.*, $D = [D_{enc}, D_{dec}]$. Same as the generator, the backbone of the discriminator is also an encoder-decoder network where face reconstruction is D_r, aiming to increase the divergence of the auto-encoder loss distributions between real and synthesised samples. The code layer of the auto-encoder is followed by D_a, D_c and D_d where $D_a(x)$ is for real-fake classification, $D_c(x)$ is for pose regression and D_d is for identity prediction. In Algorithm 1, we summarise the learning procedure of the proposed DED-GAN model. We use the Adam optimiser [68] for network training.

Algorithm 1 The DED-GAN Training Algorithm

```
Input: Training dataset $X$ and label $Y$. $X = \{x_1, x_2, \ldots, x_N\}$. $Y$ includes the pose label and identity label: $Y = \{(y^{pos}_1, y^{id}_1), (y^{pos}_2, y^{id}_2), \ldots, (y^{pos}_N, y^{id}_N)\}$. Initialise all the parameters $\theta = \{\theta_g, \theta_d\}$ in generator and discriminator, trade-off hyper-parameters $\lambda_a, \lambda_d, \lambda_c, \lambda_r, \mu_a, \mu_d, \mu_c, \mu_r$ and Adam hyper-parameter $\alpha$. The number of iteration t← 0.
Output: $\theta = \{\theta_g, \theta_d\}$
1: while $\theta_g$ does not converge do.
2: t← t+1.
3: Sample noisy data $Z$ and pose code $C$ and compute the cost of $L^t(D)$ by $L^t(D) \leftarrow \lambda_a L^{D^t}_{adv}(X) + \lambda_d L^{D^t}_{id}(X) + \lambda_c L^{D^t}_{pos}(X) + \lambda_r L^{D^t}_{pixel}(X, Z, C)$ using equations (5)-(9).
4: Compute the back propagation error to optimise discriminator $\Theta^t_d \leftarrow Adam(\nabla_{\theta^t_d} L^t(D), \alpha)$.
5: Sample noisy data $Z$ and pose code $C$ and generate data $X_g = \{G(x_1, z_1, c_1), G(x_2, z_2, c_2), \ldots, G(x_N, z_N, c_N)\}$.
6: Compute the cost of $L^t(G)$ by $L^t(G) \leftarrow \mu_a L^{G^t}_{adv}(X) + \mu_d L^{G^t}_{id}(X) + \mu_c L^{G^t}_{pos}(X) + \mu_r L^{G^t}_{pixel}(X, Z, C)$ using equations (10)-(14).
7: Fix the discriminator parameter $\Theta^t_d$ and compute the back propagation error to optimise generator $\Theta^t_g \leftarrow Adam(\nabla_{\theta^t_g} L^t(G), \alpha)$.
8: end while
```

All the experiments were performed with the following settings. All face images were aligned to a canonical view of 100 × 100 in size. Randomly sampled regions of size 96 × 96 pixels selected from 96 × 96 each aligned face were cropped for data augmentation. The image intensity was linearly scaled to the range of [-1,1]. All weights in the networks were initialized by a normal distribution with 0 mean and standard deviation of 0.02. We set the diversity ratio, β, to 0.9. $k_t \in [0, 1]$ controls how much emphasis is put on $L(G(x))$

TABLE 1. DED-GAN and its partial variants performance comparison.

Model	0°	±15°	±30°	±45°	±60°	Average
DED-GAN(-D^c)	99.62	98.20	95.78	92.04	86.11	93.47
DED-GAN(-D^r)	99.33	98.62	96.86	92.39	86.20	93.92
DED-GAN(-D^a)	99.48	99.04	97.47	93.47	85.65	94.36
DED-GAN(using pose classification)	99.72	99.15	97.76	94.12	84.96	94.64
DED-GAN*	**99.95**	**99.45**	**98.02**	**94.88**	**87.82**	**95.75**

during the network optimisation. We initialise $k_0 = 0$ and update k in each training step. λ_k is the learning rate for k. We set λ_k to 0.001 in our experiments. We define the trade-off between the respective components of the loss function by setting $\lambda_a = 1$, $\lambda_d = 1$, $\lambda_c = 0.1$, $\lambda_r = 10$, $\mu_a = 1$, $\mu_d = 1$, $\mu_c = 0.1$ and $\mu_r = 10$ through numerous experiments. All experiments were run on a NVIDIA GeForce GTX Titan Xp card with CUDA 8.0 and cuDNN 6.0, implemented in Pytorch.

V. EXPERIMENTS

A. EXPERIMENTAL SETTINGS AND DATASETS

We evaluate DED-GAN qualitatively and quantitatively under both constrained and unconstrained scenarios for face synthesis across poses and PIFR. Our models were trained separately on the Multi-PIE [69] and CASIA [67] datasets. For the qualitative evaluation, we show visualised results of face synthesis on Multi-PIE, CASIA and CFP [70]. For the quantitative evaluation, we measure face recognition performance using the learned facial representations with a cosine distance metric on the Multi-PIE, CFP and LFW [71] datasets.

The **Multi-PIE** database is the largest multi-view face recognition benchmark in the constrained scenario. It contains more than 750,000 images of 337 identities recorded in five months. Each identity has images captured under 15 poses and 20 illuminations. These images were captured in four sessions during different periods. Like the previous methods, we evaluate our algorithm on a subset of the Multi-PIE database, where each identity has images from all the four sessions under nine poses from yaw angles −60° to +60°. For a fair comparison, we follow the setting used in DR-GAN [18]. We evaluate our method on the Multi-PIE dataset setting 2. The first 200 subjects are used for training and the remaining 137 subjects are used for testing. Different from DR-GAN in which the supervised pose information is used, we use MTCNN to extract five landmarks and then transform the landmarks to a pose label. In testing, one frontal view with neural illumination is used as the gallery image and other images are used as probes. Therefore, we have $N^d = 200$ for identity classification, $N^p = 1$ for pose regression, $N^a = 1$ for real/fake classification and $N^r = 3\times96\times96$ for colour image reconstruction. We set the dimension of the embedding feature and uncompressed noise to $N^f = 320$ and $N^z = 50$ respectively.

The **CASIA** database offers 494,414 in-the-wild face images of 10,575 subjects. It is a widely used large-scale database for face recognition. We train our model on this dataset to evaluate the performance of our model on a realistic dataset. We have $N^d = 10,575$, $N^p = 1$. N^f and N^z are set as for Multi-PIE. We also evaluate the performance of our model in terms of the quality of synthesised face poses.

The **CFP** database contains 7,000 images of 500 subjects, where each subject has 10 frontal and 4 profile face images. The data are randomly organized into 10 splits, each containing an equal number of frontal to frontal and frontal to profile pairs, with 350 intra pairs and 350 non-matching pairs, respectively. We evaluate the face verification performance in terms of front-to-front and profile-to-front matching. We also evaluate the performance of our model on its ability to synthesise faces across pose variations.

The **LFW** database contains 13,233 face images of 5,749 identities. The images were obtained by trawling the internet followed by face centering, scaling, and cropping based on the bounding boxes provided by an automatic face detector. The LFW data have large in-the-wild variability, *e.g.*, in-plane rotations, non-frontal poses, non-frontal illumination, varying expressions and so on. The verification set consists of 10 folders, each with 300 matching pairs and 300 non-matching pairs. We measure the face verification performance and compare it with existing methods.

B. ABLATION STUDY

Our discriminator is designed as a multi-task CNN with four components, namely D^a, D^c, D^d and D^r for real/fake classification, pose regression, identification and face reconstruction respectively. While D^d surely plays a significant role in assisting the model to preserve the face identity, it is instructive to understand the role of the remaining components. In this subsection, the effect of the four loss functions on the recognition performance is investigated. The results are presented in Tab. 1 which reports the recognition performance of DED-GAN partial variants with each of D components removed. While the variant without adversarial loss D^a exhibits a slight performance drop, the models without face reconstruction D^r and pose regression D^c losses are degraded more severely. When removing D^c, there is no pose label to supervise the face discrimination, especially for the profile faces. The average accuracy of DED-GAN partial variants without pose estimation reduces from 95.75% to 93.47%. This can be attributed to the pose information being entangled with identity in the feature representation.

Tab. 1 also presents the performance of our model without face reconstruction D^r. The average accuracy drops

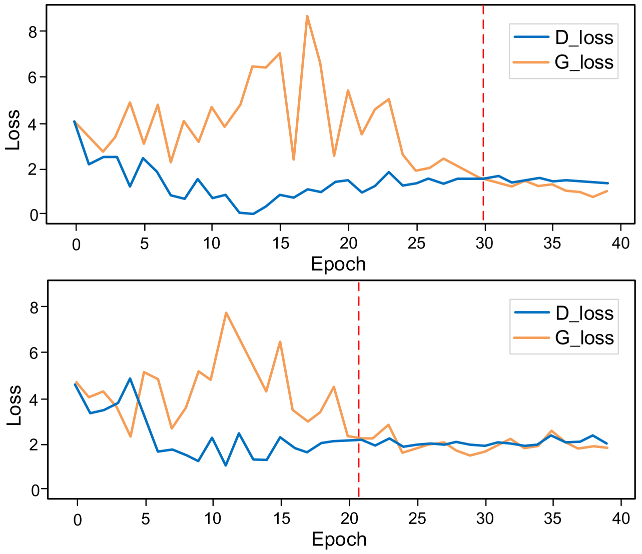

FIGURE 2. **Comparison of training loss of DED-GAN without (top) and with (bottom) pixel-wise loss on Multi-PIE (The top shows that training losses of generator and discriminator of DED-GAN without pixel-wise loss (DED-GAN ($-D^r$)); The bottom shows that training losses of generator and discriminator of DED-GAN with pixel-wise loss (DED-GAN*)).**

from 95.75% to 93.92%. This shows that facial reconstruction is almost equally important to pose estimation. This suggests that the encoder-decoder structured discriminator successfully balances the training of the two players in GAN.

To gauge the impact of using pose regression, rather than pose classification, we train separate DED-GAN models using the respective formulations. The results show that the performance of the model based on pose classification is lower by about 1%. Thus continuous pose variation used for regression benefits for preserving more information about the pose.

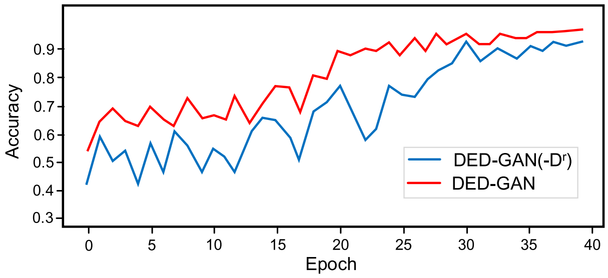

FIGURE 3. **Comparison of test accuracy of DED-GAN without (blue) and with (red) pixel-wise loss on Multi-PIE.**

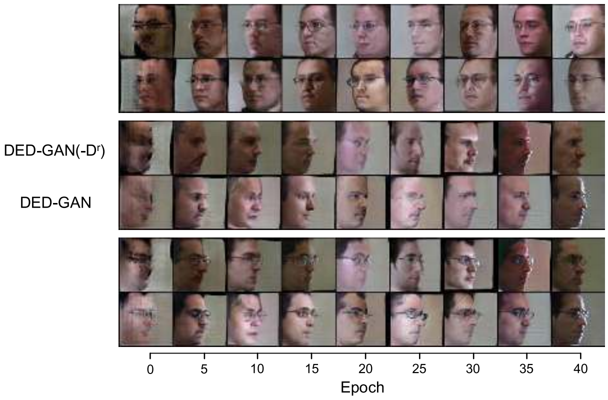

FIGURE 4. **Comparison of some synthesised faces of DED-GAN without (top) and with (bottom) pixel-wise loss on Multi-PIE.**

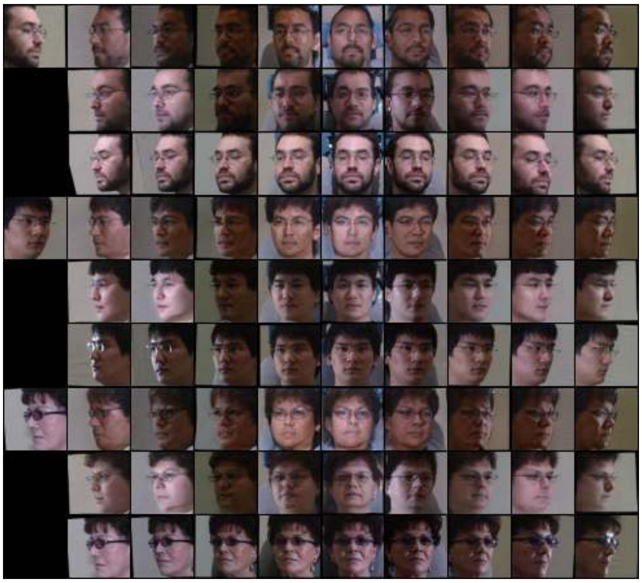

FIGURE 5. **Comparison of DR-GAN and DED-GAN generated images on Multi-PIE. Given three input images (the left column), the first, fourth and seventh rows shows the faces synthesised by the DR-GAN; the second, fifth and eighth rows show the faces synthesised by DED-GAN; the third, sixth and ninth rows show the ground truth of nine poses within the degree from −60° to 60°.**

The pixel-wise loss could effectively balance the generator and discriminator and get a fast convergence of training. To evaluate whether the pixel-wise loss could boost the convergence performance of DED-GAN, we compare the GAN loss with and without reconstruction task. Fig. 2 shows that DED-GAN without pixel-wise loss almost achieves convergence after 30 epochs. However, DED-GAN with pixel-wise loss gets a balance between generator and discriminator after about 20 epochs. The additional reconstruction task with pixel-wise loss suggests a fast and stable training manner between the generator and the discriminator of GAN. We also compare the performance of DED-GAN with and without pixel-wise loss on the test accuracy and synthesised faces. As shown in Fig. 3, the DED-GAN with pixel-wise loss almost gets a stable test accuracy after 20 epochs training, while the DED-GAN without pixel-wise loss gets a stable accuracy at about 30 epochs. Fig. 4 shows the synthesised faces of DED-GAN with and without pixel-wise loss every five epochs during training. The result also shows that DED-GAN with pixel-wise loss could boost the quality of synthesised faces during training.

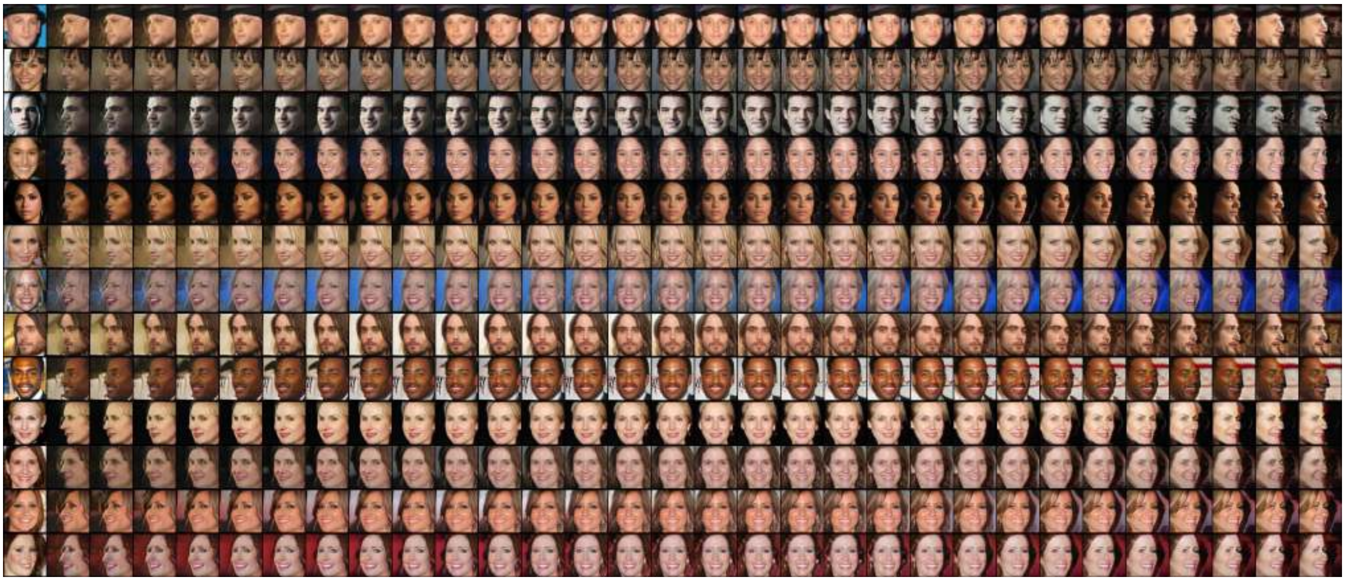

FIGURE 6. Face manifold across poses on the CASIA database (Input is at the first column, the faces from the 2nd column to the last one are the manifold of synthesised faces with the same identity by changing the value of pose code from −17 to 17).

C. FACE SYNTHESIS

To verify the performance of our method in terms of the quality of face synthesis across poses, several experiments are conducted on Multi-PIE, CASIA and CFP datasets. In the first experiment, we compared the synthesised faces with different poses between DR-GAN and our method on Multi-PIE. The synthesised faces are verified on the test set of the Setting 2. Hence, there is no overlap of subjects between the training and test datasets. Given a random input face, we generate synthesised faces within a pose range of ±60°. The experimental results are shown in Fig. 5. We can see that the pose estimation capability helps to generate faces across poses and successfully disentangle pose variation from the feature vector in both methods. However, the quality of the faces synthesised by our method appears to be better than that of those output by the DR-GAN in texture, shape, as well as identity preserving characteristics.

For an objective evaluation of the relative quality of faces generated by the two types of GANs, we use the Fréchet Inception Distance (FID) [72]. For a feature function ϕ (by default, the Inception network's convolutional feature), FID models $\phi(p_d)$ and $\phi(p_g)$ as Gaussian random variables with empirical means μ_d, μ_g and empirical covariance Σ_d, Σ_g. FID is expressed as $FID(p_d, p_g) = ||\mu_d - \mu_g|| + Tr(\Sigma_d + \Sigma_g - 2(\Sigma_d \Sigma_g)^{1/2})$, which is the Fréchet distance between the two Gaussian distributions. Tab. 2 compares the FID scores between DR-GAN and DED-GAN. DED-GAN achieves a lower FID score than DR-GAN, which means that the faces synthesised by DED-GAN are more similar to real ones than those produced by DR-GAN.

TABLE 2. Comparison of FID score.

Model	FID score
DR-GAN	71.25
DED-GAN	**57.03**

To further demonstrate the ability to disentangle the pose generative factor from other face attributes, we also evaluate the performance of our model on face synthesis across poses on another two uncontrolled datasets CASIA and CFP. We use MTCNN to extract five facial landmarks for each face and then transform the landmarks to pose label by a statistical shape model. The CASIA facial distribution across poses is illustrated in Fig. 7 where the value zero denotes the frontal face. Note that, different from the previous methods, DED-GAN can rotate an input face to any pose controlled explicitly by the pose code. Hence, DED-GAN can synthesise both frontal and profile faces. Fig. 6 shows the pose manifold of generated faces by changing the value of the pose code. Every row denotes the faces with the same identity. The first column is the input face and the other columns show the manifold of the synthesised faces with a smoothly changing value of the pose code from -17 to 17. We can see that our model preserves the identity well as we change the pose code. It also shows that the pose variation is explicitly untangled from the other face attributes including identity.

We also test the face frontalisation performance for unseen faces on the CFP dataset as shown in Fig. 8. Every column shows the faces of the same identity. Given an input profile face, we separately generate the frontal faces by DR-GAN and our method. The up and down rows show the input profile faces and paired real frontal faces separately. The second and third rows show the synthesised frontal faces by setting the pose code to zero. We can see that both methods can untangle the face representation from pose variation and generate frontal faces. However, the faces synthesised by our method appear better in terms of texture detail and in preserving the face identity.

TABLE 3. Rank-1 recognition rates (%) across views, illuminations and sessions under Multi-PIE.

Model	0°	±15°	±30°	±45°	±60°	Average
Zhu *et al.* [73]	95.70	92.80	83.70	72.90	60.10	79.30
Yim *et al.* [10]	99.50	95.00	88.50	79.90	61.90	83.30
DR-GAN [51]	97.00	94.00	90.10	86.20	83.20	89.20
DR-GAN$_{am}$ [18]	98.10	95.00	91.30	88.00	85.80	90.80
FF-GAN [74]	-	94.60	92.50	89.70	85.20	-
Light CNN [75]	-	98.59	97.38	92.13	62.09	-
DED-GAN*	**99.95**	**99.45**	**98.02**	**94.88**	**87.82**	**95.75**

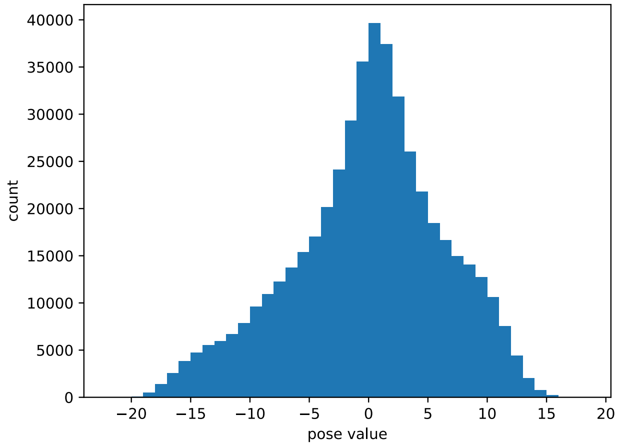

FIGURE 7. Face distribution across poses on the CASIA database.

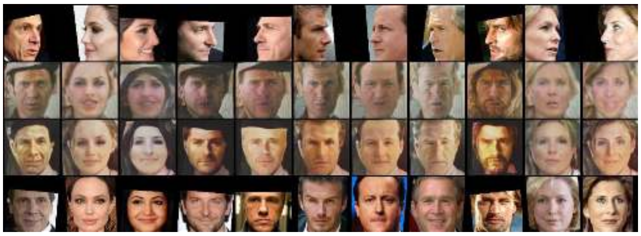

FIGURE 8. Some face frontalisation results comparison on CFP database (from top to bottom: input images, DR-GAN frontalised faces, our frontalised faces, real frontal faces).

D. FACE RECOGNITION

One motivation for disentangled face representation learning is to see, whether the untangled representation helps to preserve the identity information, and thus boost the performance in face recognition. To verify this, we also show quantitative results obtained in PIFR experiments. We evaluate our method on Multi-PIE, CFP and LFW for identification and verification tasks. The features are extracted from G_{enc} in all the experiments. The cosine distance between two representations is used for face recognition in the test step.

1) FACE IDENTIFICATION ON THE MULTI-PIE DATABASE

In the first experiment in PIFR, we evaluate the performance of DED-GAN on the Multi-PIE dataset. We compare our method with other state-of-the-art face recognition methods. Our model achieves the best accuracy in different pose categories, with the most significant improvement noted for the profile faces as shown in Tab. 3. It shows that our method can remove the effects of the pose and retain the intrinsic face shape and structure information of identity.

TABLE 4. Face verification accuracy(%) comparison on on CFP.

Model	Frontal-Frontal	Frontal-Profile
Sengupta *et al.* [70]	96.40 ± 0.69	84.91 ± 1.82
Sankarana *et al.* [76]	96.93 ± 0.61	89.17 ± 2.35
DR-GAN [51]	97.13 ± 0.62	90.82 ± 0.28
Human	96.24 ± 0.67	94.57 ± 1.10
DED-GAN*	**97.99±0.85**	**91.58±1.38**

2) FACE VERIFICATION ON THE CFP DATABASE

To further demonstrate the advantages of our method in PIFR, we evaluate it on an uncontrolled dataset. For the in-the-wild setting, we train our model on CASIA and test it on the CFP database. The experiments performed on the CFP dataset aim to compare the capacity of the face verification approaches across diverse poses. More specifically, the matching is performed between the frontal view (yaw angle $< 10°$) and profile view (yaw angle $> 60°$). The evaluation reports the mean and standard deviation of accuracy, over 10 splits, for both frontal to frontal and frontal to profile face verification settings. The verification results are shown in Tab. 4. Our method again yields better verification performance on both frontal-frontal and frontal-profile matching sub-tasks. Thanks to the more stable training structure and more detailed pose information injected into our method, DED-GAN achieves about a one percent performance improvement over DR-GAN.

3) FACE VERIFICATION ON THE LFW DATABASE

To evaluate the performance on the in-the-wild dataset further, we test the models described in the previous subsection on the LFW database. Tab. 5 shows the accuracy achieved by different methods. As expected, our method DED-GAN delivers the best accuracy, namely 97.52%, which is comparable with other state-of-the-art methods. Although DED-GAN is not trained on the LFW dataset, the untangled discriminative representation generalises to other datasets, including in-the-wild datasets.

TABLE 5. Face verification accuracy (%) comparison on LFW.

Model	Accuracy (%)
LFW-3D [8]	93.62
LFW-HPEN [77]	96.25
FF-GAN [74]	96.42
DED-GAN*	**97.52**

VI. CONCLUSION

We propose a new GAN-based model (DED-GAN) for disentangled representation learning to address the challenging problem of pose-invariant face recognition and photo-realistic face synthesis across poses. To the best of our knowledge, this is the first time that a dual encoder-decoder structured GAN has been used to learn disentangled face representation. The encoder-decoder structured generator is used for face rotation and learning disentangled face representation. The encoder-decoder structured discriminator is used for facial reconstruction and for predicting identity, as well as for estimating the pose. The Encoder-decoder structured discriminator with the additional pixel-wise loss improves the training efficiency and stability of our GAN. A continuous pose encoding provides more detail pose information and benefits the discriminative representation by untangling the identity and pose. Extensive quantitative and qualitative experimental results show that our method is competitive compared to state-of-the-art approaches to PIFR and to face synthesis across poses. In the future, we plan to incorporate more discriminative information into the design of DED-GAN by extending the network to deal explicitly with other image generative factors, including illumination, expression, age and occlusion.

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[2] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.

[3] J. Kong, M. Chen, M. Jiang, J. Sun, and J. Hou, "Face recognition based on CSGF $(2D)^2$ PCANET," *IEEE Access*, vol. 6, pp. 45153–45165, 2018.

[4] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales, "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 142–150.

[5] X. Song, Z.-H. Feng, G. Hu, J. Kittler, and X.-J. Wu, "Dictionary integration using 3D morphable face models for pose-invariant collaborative-representation-based classification," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2734–2745, Nov. 2018.

[6] W. Wang, X. Chen, S. Zheng, and H. Li, "Fast head pose estimation via rotation-adaptive facial landmark detection for video edge computation," *IEEE Access*, vol. 8, pp. 45023–45032, 2020.

[7] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1883–1890.

[8] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304.

[9] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3871–3879.

[10] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 676–684.

[11] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3D morphable face models and their applications," in *Proc. Int. Conf. Articulated Motion Deformable Objects*. Berlin, Germany: Springer, 2016, pp. 185–206.

[12] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin, "Gaussian mixture 3D morphable face model," *Pattern Recognit.*, vol. 74, pp. 617–628, Feb. 2018.

[13] J.-S. Chan, G.-S.-J. Hsu, H.-C. Shie, and Y.-X. Chen, "Face recognition by facial attribute assisted network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3825–3829.

[14] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4838–4846.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[16] Z.-H. Feng, J. Kittler, W. Christmas, and X.-J. Wu, "A unified tensor-based active appearance face model," 2016, *arXiv:1612.09548*. [Online]. Available: http://arxiv.org/abs/1612.09548

[17] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1623–1632.

[18] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3007–3021, Dec. 2019.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[20] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain Sci.*, vol. 40, pp. 1–72, Nov. 2017.

[21] C. Hu, X.-J. Wu, and J. Kittler, "Semi-supervised learning based on GAN with mean and variance feature matching," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 4, pp. 539–547, Dec. 2019.

[22] C. Hu, X.-J. Wu, and Z.-Q. Shu, "Discriminative feature learning via sparse autoencoders with label consistency constraints," *Neural Process. Lett.*, vol. 50, pp. 1079–1091, Aug. 2018.

[23] K. Ridgeway, "A survey of inductive biases for factorial representation-learning," 2016, *arXiv:1612.05299*. [Online]. Available: http://arxiv.org/abs/1612.05299

[24] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P Burgess, M. Bosnjak, M. Shanahan, M. Botvinick, D. Hassabis, and A. Lerchner, "SCAN: Learning hierarchical compositional visual concepts," 2017, *arXiv:1707.03389*. [Online]. Available: http://arxiv.org/abs/1707.03389

[25] Z. Tang, J. Yang, Z. Pei, X. Song, and B. Ge, "Multi-process training GAN for identity-preserving face synthesis," *IEEE Access*, vol. 7, pp. 97641–97652, 2019.

[26] X. Luan, H. Geng, L. Liu, W. Li, Y. Zhao, and M. Ren, "Geometry structure preserving based GAN for multi-pose face frontalization and recognition," *IEEE Access*, vol. 8, pp. 104676–104687, 2020.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, vol. 3, 2017, pp. 1–13.

[30] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2539–2547.

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[32] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*. [Online]. Available: http://arxiv.org/abs/1609.03126

[33] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*. [Online]. Available: http://arxiv.org/abs/1703.10717

[34] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[36] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: http://arxiv.org/abs/1710.10196

[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[39] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[40] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 1, no. 2, p. 3.

[41] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[42] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project Stanford CS231N, Convolutional Neural Netw. Vis. Recognit., Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[43] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional CycleGAN," 2017, *arXiv:1705.09966*. [Online]. Available: https://arxiv.org/abs/1705.09966

[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[45] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[46] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.

[47] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4030–4038.

[48] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2107–2116.

[49] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[50] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view image generation from a single-view," 2017, *arXiv:1704.04886*. [Online]. Available: http://arxiv.org/abs/1704.04886

[51] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 3, no. 6, p. 7.

[52] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2089–2093.

[53] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," 2017, *arXiv:1703.01560*. [Online]. Available: http://arxiv.org/abs/1703.01560

[54] C. H. Chan, M. A. Tahir, J. Kittler, and M. Pietikainen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1164–1177, May 2013.

[55] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 7, pp. 1160–1169, 1985.

[56] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[58] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.

[59] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[60] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 315–323.

[61] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 499–515.

[62] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[63] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2892–2900.

[64] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*. [Online]. Available: http://arxiv.org/abs/1502.00873

[65] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[66] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[67] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[69] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[70] 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 *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[71] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images, Detection, Alignment, Recognit.*, 2008, pp. 1–15.

[72] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local NASH equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[73] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 217–225.

[74] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–10.

[75] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[76] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–8.

[77] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.

CONG HU received the Ph.D. degree from Jiangnan University, Wuxi, China, in 2019. He was a Visiting Ph.D. Student with the Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., from July 2018 to July 2019. He is currently a Lecturer with the School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include pattern recognition, deep learning, and computer vision.

ZHENHUA FENG (Member, IEEE) received the Ph.D. degree from the Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2016. He is currently a Lecturer with the Department of Computer Science, University of Surrey. He has published more than 40 scientific articles in top-ranking conferences and journals, such as CVPR, ICCV, IJCAI, IJCV, TIP, TIFS, TCSVT, and so on. His research interests include pattern recognition, machine learning, and computer vision. He received the European Biometrics Industry Award from the European Association for Biometrics, in 2017.

XIAOJUN WU (Member, IEEE) received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively. He is currently a Professor of artificial intelligent and pattern recognition with Jiangnan University, Wuxi, China. His current research interests include pattern recognition, computer vision, fuzzy systems, neural networks, and intelligent systems.

JOSEF KITTLER (Life Member, IEEE) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively.

He is currently a Distinguished Professor of machine intelligence with the Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He has published the textbook *Pattern Recognition: A Statistical Approach* and over 700 scientific articles. His publications have been cited more than 68,000 times (Google Scholar). He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He is a Series Editor of Springer Lecture Notes on Computer Science. He also serves on the editorial boards for *Pattern Recognition Letters*, *Pattern Recognition and Artificial Intelligence*, and *Pattern Analysis and Applications*. He has served as a member of the Editorial Board for the IEEE Transactions on Pattern Analysis and Machine Intelligence, from 1982 to 1985. He has served on the Governing Board for *International Association for Pattern Recognition* (IAPR) as one of the two British representatives, from 1982 to 2005, and the President of the IAPR, from 1994 to 1996.