FA-GAN: Face Augmentation GAN for Deformation-Invariant Face Recognition

Mandi Luo, Jie Cao, Xin Ma, Xiaoyu Zhang, and Ran He, *Senior Member, IEEE*

***Abstract*— Substantial improvements have been achieved in the field of face recognition due to the successful application of deep neural networks. However, existing methods are sensitive to both the quality and quantity of the training data. Despite the availability of large-scale datasets, the long tail data distribution induces strong biases in model learning. In this paper, we present a Face Augmentation Generative Adversarial Network (FA-GAN) to reduce the influence of imbalanced deformation attribute distributions. We propose to decouple these attributes from the identity representation with a novel hierarchical disentanglement module. Moreover, Graph Convolutional Networks (GCNs) are applied to recover geometric information by exploring the interrelations among local regions to guarantee the preservation of identities in face data augmentation. Extensive experiments on face reconstruction, face manipulation, and face recognition demonstrate the effectiveness and generalization ability of the proposed method.**

***Index Terms*— Face augmentation, deformation-invariant face recognition, face disentanglement, graph convolutional networks.**

I. Introduction

FACE recognition has been an active area over the past decades and has shown extremely high value in both research and practical applications. The notable success of deep learning has enabled significant progress in face recognition [1]–[7], [7]–[17]. However, face deformation [18], i.e. face variations caused by pose, expression, and other facial movements and evolution, severely interferes with face recognition performance under in-the-wild conditions, making it a long-standing challenge.

Recent research substantiates that utilizing synthesized faces from generation models [23]–[25] assists recognition models in reducing the influences from face deformations. Given an arbitrary face, the corresponding normalized face [26], i.e., the face whose identity is preserved while other factors are normalized, can be inferred by a generation model. Then, a recognition model extracts the identity representation [27] from the normalized face or both the normalized and the original faces for the subsequent verification/recognition task. Such *recognition via generation* [28], [29] mitigates the issues caused by face deformations, such as pose, expression, and other factors.

Following this line of study, existing works have made exceptional contributions in constrained environments. Their models have achieved both satisfying visual quality and high-level face recognition rates on some benchmark datasets, such as Multi-PIE [30] and M^2FPA [31], which are collected under constrained environments. However, there still exist some ongoing issues when these models are applied to real-life situations. Initially, the approaches are mainly based on the 3D Morphable Model (3DMM) [32]. Researchers utilize 3DMM to reconstruct 3D face models and then render the corresponding normalized faces. 3D-based methods [33], [34] have difficulties in completing occluded facial regions and dealing with extreme cases. Current methods [35], [36] based on Generative Adversarial Network dominate the task of *recognition via generation* and have achieved notable results. However, these approaches still face challenges when applied to real-world scenarios.

We note that collecting larger datasets can mitigate this issue as well. However, this is usually infeasible since the cost is prohibitively high in practice. Labeling data collected from real-life situations is even more expensive, and many samples are difficult to lable. On the other hand, as shown in Fig. 1, these data always fall into unbalanced distributions in terms of face deformations such as large pose. These facts lead to overfitting problems.

To eliminate the influence of the long tail distribution of the data, we propose a Face Augmentation Generative Adversarial Network (FA-GAN) to augment the existing datasets by generating faces with various deformations. We focus on identity preservation and regard other aspects of the input image, such as background, as nuisance factors. Moreover, to avoid fuzziness in the generated face images, especially in dealing with extreme deformations, we choose to handle the geometry and texture information separately. Our proposed FA-GAN is composed of two branches. The first branch is a graph-based Geometric Preserving Module (GPM) for learning

Manuscript received August 4, 2020; revised December 8, 2020 and January 12, 2021; accepted January 12, 2021. Date of publication January 21, 2021; date of current version February 12, 2021. This work was supported in part by the Beijing Natural Science Foundation under Grant JQ18017; in part by the National Natural Science Foundation of China under Grant 61721004, Grant U20A20223, and Grant U2003111; and in part by the Youth Innovation Promotion Association CAS under Grant Y201929. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vitomir Štruc. *(Corresponding author: Xiaoyu Zhang.)*

Mandi Luo, Jie Cao, Xin Ma, and Ran He are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: luomandi2019@ia.ac.cn; jie.cao@cripac.ia.ac.cn; xin.ma@cripac.ia.ac.cn; rhe@nlpr.ia.ac.cn).

Xiaoyu Zhang is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: zhangxiaoyu@iie.ac.cn).

Digital Object Identifier 10.1109/TIFS.2021.3053460

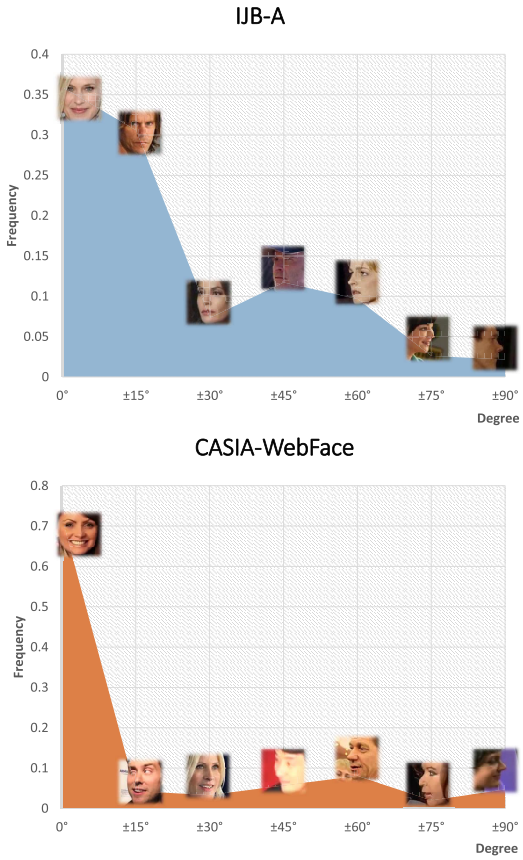


Fig. 1. Data distribution in IJB-A [19] and CASIA-WebFace [20] in terms of different poses. Note that there are no official labels for the different poses with accurate degrees. We detect face landmarks using [21] and estimate the degree based on the distance between the left eye center and right eye center. We assume that the possible errors lie in an acceptable margin. Note that the face images approximately obey the long tail distribution [22].

the geometry information. The second branch is the Face Disentanglement Module (FDM), which is a novel two-stage disentanglement module responsible for disentangling deformation attributes representations from identity representations. On one hand, the two branches deal with high-level and low-level information separately, making them more focused compared with directly image-to-image translation. On the other hand, benefiting from the end-to-end training strategy, the two branches are designed to coordinate with each other to obtain mutual improvements.

The preservation of face geometry such as face shape and features is of great importance for face recognition. However, while generating normalized faces, previous schemes [28], [36] use mean facial landmarks calculated on training datasets. These schemes cause geometry distortion in the synthesized faces because they neglect the fact that landmarks are identity-dependent. To address this issue, we estimate the normalized face parsing map $\mathbf{I}_p$ for each input by learning identity-dependent geometric information. Since the face parsing map contains both semantic information and the spatial distribution of different face regions, it is able to perform as a suitable guidance with high-level information for image generation. Moreover, considering that human faces are non-rigid objects, the process of face deformation variations such as face rotation and expression changes in fact belong to nonlinear transformation. As demonstrated in previous literature [37], [38], Graph Convolutional Networks (GCNs) [39] are more suitable than Convolutional Neural Networks (CNNs) in dealing with non-Euclidean relations. Specifically, we treat different face parts, which are divided approximately based on semantic meaning, as different nodes of a graph. Then we apply GCNs to learn the interrelations and high-level similarities between these nodes under the supervision of $\mathbf{I}_p$, and the edges between the nodes are updated accordingly during the training phase. In this way, our model is able to jointly explore the spatial and semantic relations of different face regions, thus the identity-dependent geometry information can be well preserved. The learned $\hat{\mathbf{I}}_p$ is utilized for guiding the generation process, which is conducted by the second branch, called the FDM.

Information disentanglement has been widely used in face generation tasks [35], [40]. Unfortunately, when large deformations are taken into account, existing approaches still struggle between learning useful representations and maintaining visual quality. This may occur because existing approaches usually decouple representations simultaneously. However, we notice that some deformations, such as expression, are correlated with identity. As depicted in Fig. 3, there are intersections between identity representations and deformation representations. Naturally, we propose the idea of hierarchical representation disentanglement. The proposed FDM branch is designed with two stages. The first stage is used for disentangling identities from nuisance factors, which are termed "noise". Furthermore, the second stage disentangles the deformation attribute representations. With these representations well disentangled, we are able to achieve face manipulation based on actual needs. Then, we can augment real-life face datasets to finetune recognition models and boost their recognition/verification performance.

Extensive qualitative and quantitative results have been obtained on multiple datasets, including Multi-PIE [30], M^{2}FPA [31], LFW [41], IJB-A [19], and MegaFace [42]. We thoroughly evaluate the recognition/verification accuracy [20] of numerous well-known recognition models [8], [11], [13], [43]–[47] on the augmented CASIA-WebFace. The performance improvements demonstrate that our proposed method represents the identity and deformations well not only in constrained environments but also in unconstrained environments. This proves that our model trained on constrained datasets can be adapted to unconstrained datasets without extra efforts, which shows the effectiveness and generalization of our proposed FA-GAN.

Our contributions are summarized as follows:

- We propose a novel hierarchical disentangled representation learning scheme, namely Face Augmentation Generative Adversarial Network (FA-GAN), for identity-preserving face synthesis with various deformations, such as large poses.
- We present the Geometry Preserving Module (GPM) to extract geometric information by exploring both spatial and semantic relations among different face regions. Preserving identity-related geometric information provides

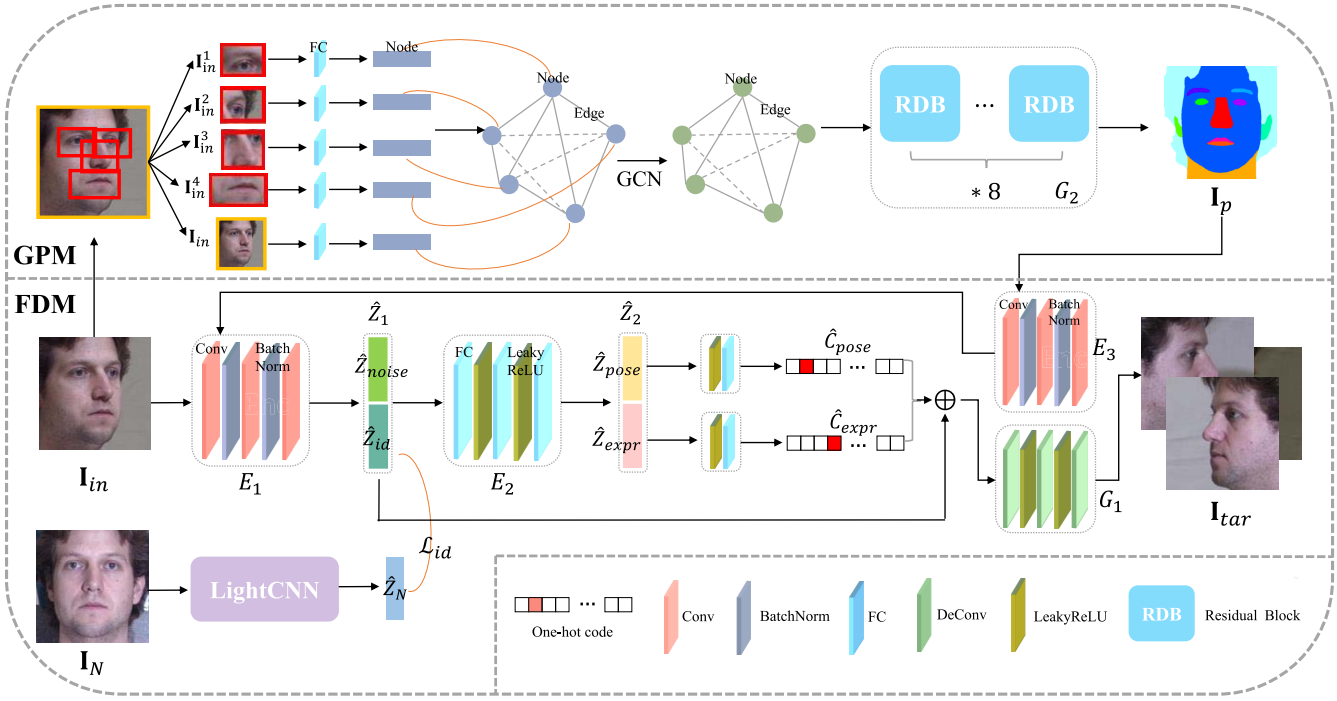


Fig. 2. Overall architecture of our proposed FA-GAN. The FA-GAN contains two branches, which address geometry preservation and face feature disentanglement, respectively. The first branch, called GPM, models the input face $\mathbf{I}_{in}$ as graph $G = \{V, E\}$ by treating the latent embeddings of different face regions as nodes $v \in V$ and their interrelations as $e \in E$. The GPM decodes the updated graph G' into the normalized face parsing map $\mathbf{I}_p$, which is used in FDM. This second branch disentangles the encoded face feature embeddings into identity representation $\hat{Z}_{id}$ and deformation codes $(\hat{C}_{pose}, \hat{C}_{pose})$ under the guidance of $\hat{\mathbf{I}}_p$. The encoded embedding triplet $< Z_{id}, C_{expr}, C_{pose} >$ can be utilized for further face manipulation and augmentation.

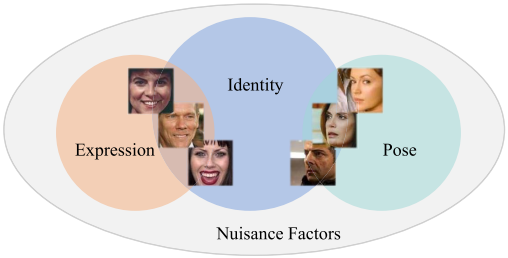


Fig. 3. Illustration of the entangled representations. The largest ellipse is the extracted representation from a certain face image. Different circles depict different attribute representations.

another advantage in both face recognition accuracy and face generation performance.

- The extensive qualitative and quantitative results prove the effectiveness of the proposed FA-GAN, especially in unconstrained environments, which are extremely challenging.

II. Related Work

A. Face Synthesis

Synthesizing desired faces is a challenging problem. Researchers have made great efforts to address this problem for years and have achieved exciting results. The earliest approaches are mostly based on classical computer graphics [48]–[50]. For instance, Zhu *et al.* [49] proposed High-Fidelity Pose and Expression Normalization (HPEN) with 3DMM [32] to recover canonical-view [51], expression-free images. Other researchers have also attempted statistical modeling. A joint frontal view reconstruction and landmark localization using a small number of frontal images were realized by solving a constrained low-rank minimization problem [52]. Later, with the introduction of GAN [23], which was a landmark event in AI history, deep learning-based methods [35], [36], [53]–[57] began occupying researchers' horizons. Goodfellow *et al.* [23] first realized synthesizing images from white noise and proved the possibility of synthesizing images with deep learning methods. Yang *et al.* [58] introduced a recurrent convolutional encoder-decoder network to capture long-term dependencies and render rotated objects. Zhu *et al.* [59] proposed cycle consistency to address circumstances where there is no paired training data. Cao *et al.* [57] introduced UV space in the training of deep learning networks to make the results more photo-realistic. Brock *et al.* [60] used orthogonal regularization in generator networks and realized balance control between the quality and diversity of the synthesized images. Karras *et al.* [61] drew inspiration from style transfer and realized detailed manipulation in face images.

Note that our proposed FA-GAN is also based on GANs. We focus on synthesizing face images with desired deformations from arbitrary deformations. Inspired by [35], we extract deformation-invariant identity representations to better preserve the identity information.

B. Representation Learning

In face recognition tasks, researchers aim at learning effective identity representations and have achieved stunning results [1]–[14], [62]–[65]. Traditionally, some works are based on sparse encoding. For instance, Zhang *et al.* [64] analyzed the working mechanism of sparse representation based classification and pointed out that the use of Collaborative Representation (CR) is efficient for face classification. Yang *et al.* [62] proposed robust sparse coding (RSC) to model the sparse coding as a sparsity-constrained robust regression problem. The RSC seeks for the maximum likelihood estimation solution of the sparse coding problem and was proved to be effective in dealing with outliers. Minaee *et al.* [63] used the scattering transform, which is a kind of convolutional network that encodes signals into multi-layer representations, to extract features from faces.

Nowadays, the methods are mainly based on deep learning networks. In 2015, Parkhi [11] proposed to realize the face recognition task with DeepFace, which is one of the most representative networks in face recognition. Liu *et al.* [45] proposed SphereFace to convert the Softmax loss from the Euclidean distance to an angular interval. Wang *et al.* [44] demonstrated that through the maximization of the cosine decision boundary, the maximum interclass difference and the minimum intraclass differences can be realized. Liu *et al.* [15] proposed the Adaptive Margin Softmax to adjust the margins for different classes adaptively, and introduced Hard Prototype Mining and Adaptive Data Sampling to make the training more effective and efficient. Cao *et al.* [7] addressed the long-tailed problem in face recognition by designing a Domain Balancing (DB) mechanism.

We adopt disentangled representation learning in this paper to perform face recognition. Disentangled representation learning [17], [66]–[71] aims to model the key factors that affect the shape of the data so that changes in a key factor only cause changes in only a certain feature of the data, while other features are not affected. Disentangled representation learning is now mainly based on deep learning. For instance, Mathieu *et al.* [72] developed a conditional generative model and separated the hidden factors of variation into complementary codes. Wang *et al.* [73] realized the re-rendering of new images with specified scene properties from a single image by proposing a Tag Disentangled GAN. InfoGAN [74] maximizes the mutual information between the latent variables and the observations to learn interpretable disentangled representations. Tran *et al.* [35] proposed a representative GAN-based disentanglement method that used labeled data to decompose representations into class-related and class-independent components. Inheriting a similar supervised manner, we develop a novel two-level disentanglement module to preserve identity information and disentangle face attributes well.

C. Graph Convolutional Network

Graph data modeling is a long-standing strategy. At first, researchers focused on statistical analysis methods [75], [76], where there was no machine learning model participating in this period. Recently, researchers have gradually shifted their attention to applying deep learning models to graph data for end-to-end modeling. There are two branches in this field, which can be categorized into spectral methods [39], [77], [78] and spatial methods [79]–[82]. In 2013, Bruna *et al.* [83] proposed the first graph convolutional neural network, where graph convolution is defined in spectral space. However, the original spectral methods are complicated in time and space, thus promoting the introduction of models in [39] and [77], which parameterize the convolution kernel to greatly reduce the time and space complexity.

We follow the design of a Graph Convolutional Network (GCN), which is an end-to-end learning method officially proposed by [39] for the field of graph embedding. In the field of computer vision, Convolutional Neural Networks (CNNs) have made great achievements mainly for Euclidean structure data. While for topological structure data, the corresponding tool is a GCN, which can handle more general structural data. GCNs utilize filters to extract the high-dimensional features of nodes and their neighborhoods in the graph and thus find the high-level similarities between nodes. GCNs have been applied to many tasks, including behavior detection [84], clustering [85], and semi-supervised learning [39]. We argue that in the field of face *recognition via generation*, a GCN is capable of extracting the representative topological structures of face images, making it more suitable than the traditional CNN.

D. Data Augmentation

It is generally accepted that larger and better datasets can boost the performance of deep learning models since they are data-driven [86], [87]. Collecting labeled datasets is time-consuming and requires high budgets, especially in fields such as face recognition where the scales of the datasets are in millions. Thus, some researchers focus on augmenting the original datasets, which is termed data augmentation.

Aiming at solving problems from the training set perspective, data augmentation strategies are based on the assumption that there is still useful information hiding in the original training dataset that can be extracted [88]. The existing methods can be mainly divided into two classes: basic image manipulations and deep learning approaches. In the basic image manipulation class, techniques such as geometric transformations [89], noise injection [90], and color space transformations [91], [92] have all made contributions. However, deep learning approaches are gradually becoming the main methods. Konno *et al.* [93] proposed "Icing on the Cake" to manipulate the modularity of neural networks. DeVries *et al.* [94] considered expanding data in the feature space. Bowles *et al.* [95] claimed that GANs present a data augmentation possibility by synthesizing samples with the appearance of real images. DA-GAN [96] is one of the methods that successfully confirms the hypothesis of [95] and achieves stunning results. Masi *et al.* [97] questioned the necessity of collecting large-scale face datasets. Kortylewski *et al.* [98] conducted face recognition experiments with fully synthetic data and demonstrated that data with strong diversity can increase the generalization across different datasets. Tang *et al.* [99] introduced the 3DMM to augment face images with desired poses.

Lv *et al.* [100] proposed five different face augmentation strategies and evaluated them respectively. Taylor *et al.* [101] presented the utilization of basic geometric data augmentation schemes, such as rotation, cropping, and so on, to augment face data. Notably, our proposed FA-GAN is also a face augmentation approach based on GANs. We realize data augmentation by synthesizing face images with desired deformations from original face images with arbitrary deformations.

III. Approach

Assume that for a given face recognition network, we choose an in-the-wild dataset $\mathcal{A} = \{I_1, I_2, I_3, \cdots\}$ for training. If we want to augment $\mathcal{A}$ for deformation-invariant face recognition, two requirements should be met. First, the identity information of a person needs to be well preserved. Second, we have to obtain disentangled deformation attribute representations, such as different poses $\{p_1, p_2, p_3, \cdots\}$ and different expressions $\{q_1, q_2, q_3, \cdots\}$, which lie in the same latent space as the identity representations. This guarantees two goals: 1) learning efficient identity representations for face recognition and 2) augmenting face datasets with customized deformation demands. Finally, we obtain the expanded dataset $\mathcal{A}' = \left\{I_1^{p_1}, I_1^{q_2}, I_2^{q_1}, I_3^{p_2}, I_3^{p_3}, \cdots\right\}$. The whole process can be realized by our proposed network with the overall architecture depicted in Fig. 2. The FA-GAN mainly contains two parts: the graph-based Geometry Preserving Module (GPM) for geometry enhancement and the Face Disentanglement Module (FDM) for representation disentanglement. In the following section, we introduce the detailed architectures and the corresponding loss functions of the proposed modules. Note that our model is trained in a supervised manner, which means paired data are required. We use the presence and absence of superscript $\hat{}$ on a variable to indicate whether it is drawn from the distribution of the generated data or the target data, respectively.

A. Geometry Preserving Module

A person's identity is substantially related to the shape and arrangement of his/her face features, which are termed face geometry features in our case. It is crucial to ensure the geometric consistency while manipulating face images. To deal with deformation variations such as large poses, previous methods [28], [36] introduced predefined average landmarks, i.e., left eye center, right eye center, nose tip, and mouth center, to guide face normalization. Since different faces do not share the same landmarks, the resulting normalized faces may suffer from topology distortions, hence degrading the preservation of identity representations.

To address this problem, we propose the geometry preserving module to estimate the normalized face parsing map $\mathbf{I}_p$ from arbitrary input face image $\mathbf{I}_{in}$. Since the normalized face parsing map contains both semantic information and the spatial distribution of different face parts of a certain person, the identity-related geometry information is well preserved. Specifically, we apply GCNs to take this responsibility. GCNs are good at determining the response of a specific node based on its neighboring nodes and are different from traditional convolutions that can only be applied to standard regular grids. We use the advantage of message passing between nodes of a GCN and utilize it for exploring the relations between different face regions. As shown in Fig. 4, we adopt a different crop strategy from that in [37]. In our case, the input image $\mathbf{I}_{in}$ is cropped into different regions, i.e., $\mathbf{I}_{in}^1$, $\mathbf{I}_{in}^2$, $\mathbf{I}_{in}^3$, and $\mathbf{I}_{in}^4$, representing the left eye, right eye, nose, and mouth regions, respectively. These regions, together with $\mathbf{I}_{in}$, are encoded into embeddings as the nodes $v \in V$ in our graph $G = \{V, E\}$. In this way, our nodes are designed with semantic meanings. The edges $e \in E$ are denoted as an adjacency matrix $\mathbf{M}$ of dimension $N \times N$, where the relations between every two nodes are taken into consideration. During training, the edges are initialized with randomly selected values and updated by back-propagation under the supervision of $\mathbf{I}_p$ until convergence. One layer of graph convolutions can be represented as:

$$\mathbf{G} = \mathbf{MXW}, \tag{1}$$

where $\mathbf{X}$ is of $N \times p$ dimension representing the input features, and $\mathbf{W}$ is the weight matrix of $p \times q$ dimension. As a result, the output $\mathbf{G}$ from one layer of graph convolution has the dimension of $N \times q$, where N is the node number and q denotes the predefined dimension of the extracted features for each node. Specifically, in our case, we stack two layers of graph convolutions and can be denoted as:

$$\mathbf{G}' = \mathbf{MXW}_1\mathbf{W}_2. \tag{2}$$

Thus, the output of the graph convolutions hidden layer is:

$$\mathbf{G}'' = \mathbf{MXW}_1. \tag{3}$$

The learned embedding features $\mathbf{G}''$ are further fed into the following generation block for face parsing map $\hat{\mathbf{I}}_p$ generation. The process is supervised by the target normalized face

Fig. 4. Different cropping strategies for graph initialization. The left denotes our cropping strategy with semantic meaning involved, and the right denotes the regular cropping strategy adopted by Li *et al.* [37].

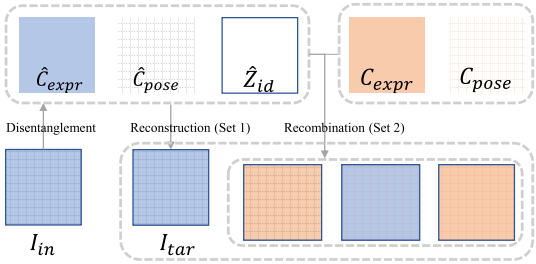


Fig. 5. Illustration of our training strategy. After the input image $\mathbf{I}_{in}$ gets disentangled, some iterations (Set 1) are executed to reconstruct $\mathbf{I}_{in}$ with the extracted $< \hat{Z}_{id}, \hat{C}_{expr}, \hat{C}_{pose} >$. Other iterations (Set 2) combine $\hat{Z}_{id}$ with target C_{expr} and C_{pose} to manipulate $\mathbf{I}_{in}$ with desired deformations.

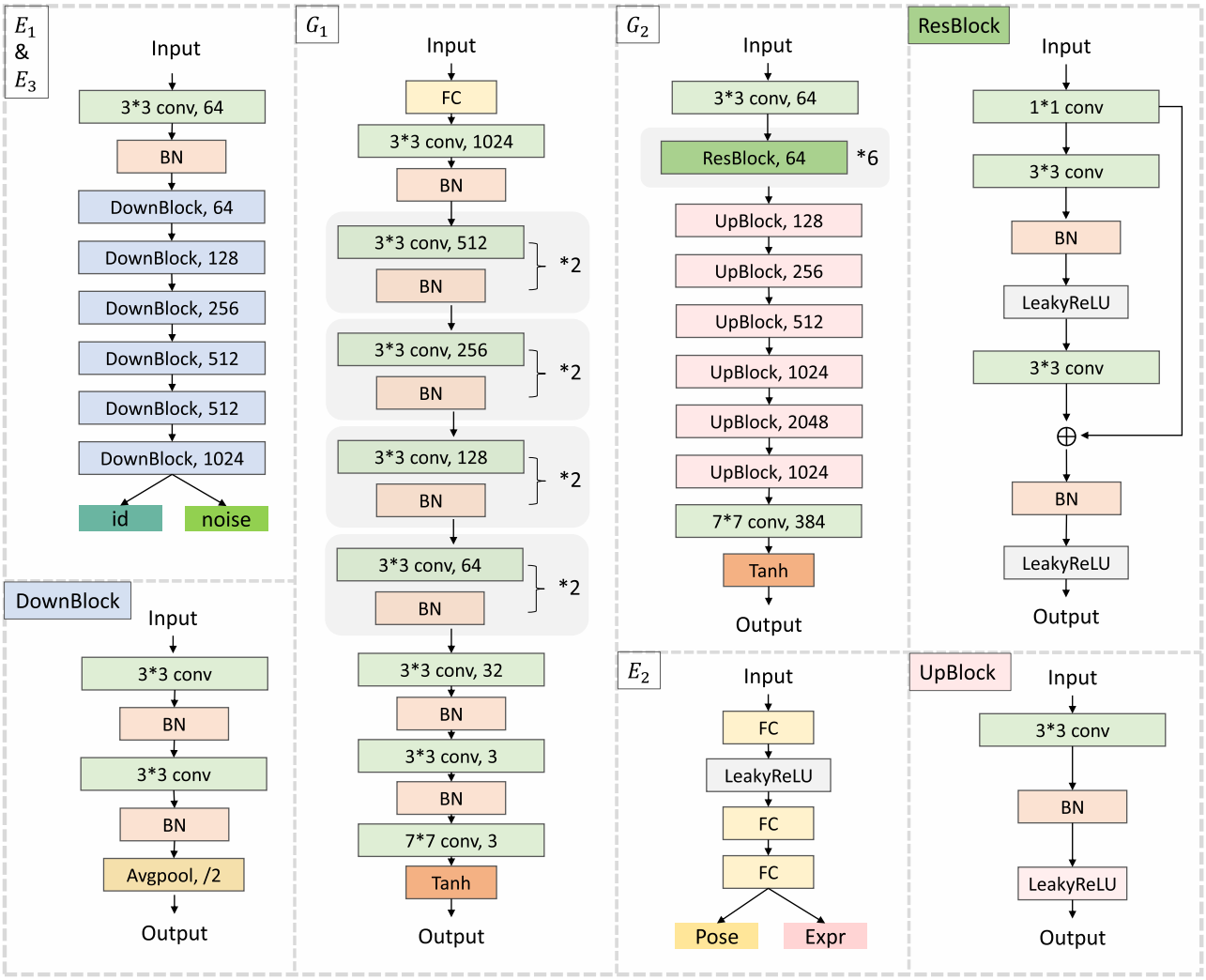


Fig. 6. Illustration of the architecture details of different subnetworks.

parsing map $\mathbf{I}_p$, which is obtained by applying the method presented in [102] to $\mathbf{I}_{in}$. The $\mathcal{L}_1$ loss is introduced that can be formulated as,

$$\mathcal{L}_p = ||\hat{\mathbf{I}}_p - \mathbf{I}_p||_1. \tag{4}$$

Represented by $\hat{\mathbf{I}}_p$, the learned relations between the nodes contain both spatial and semantic information of different face regions, depicting complete geometric information of a certain face. We further utilize $\hat{\mathbf{I}}_p$ for stabilizing geometry in the process of face manipulation and augmentation, which is discussed in the following parts.

B. Face Disentanglement Module

Disentangled feature learning has been proven to be effective in many works [35], [103], [104]. To augment face datasets with the desired deformations, we need to learn the representations of different identities and deformation attributes to further combine them freely. It is worth noting that these representations are tangled with each other in face images. Naturally, to obtain accurate representations that are as disentangled as possible, we propose the idea of hierarchical disentanglement in our proposed Face Disentanglement Module (FDM).

As shown in Fig. 2, the FDM is composed of two levels. Each of them has a different focus. The first level learns identity feature embeddings while the second level further concentrates on attribute embedding learning. Given an input image $\mathbf{I}_{in}$ with arbitrary face deformation attributes, the encoder E_1 encodes it into latent embedding $\hat{Z}_1$ under the guidance of normalized face parsing map $\hat{\mathbf{I}}_p$ generated by the first branch of the GPM. We utilize E_3 to encode $\hat{\mathbf{I}}_p$ to the same latent space as the hidden embedding $\hat{Z}_h$ which is the intermediate step in the process of obtaining $\hat{Z}_1$. Then, $\hat{\mathbf{I}}_p$ is concatenated with $\hat{Z}_h$ and fed into the remaining network of E_1, from which we obtain the encoded $\hat{Z}_1$. We then divide $\hat{Z}_1$ into two parts. The first part $\hat{Z}_{id}$ represents the identity information, and the second part $\hat{Z}_{noise}$ represents the nuisance factors. To ensure the completeness and accuracy of the extracted identity information, our network should be able to encode $\hat{Z}_{id}$ in a way similar to how it is encoded by face recognition models. The first level is responsible for this process, and

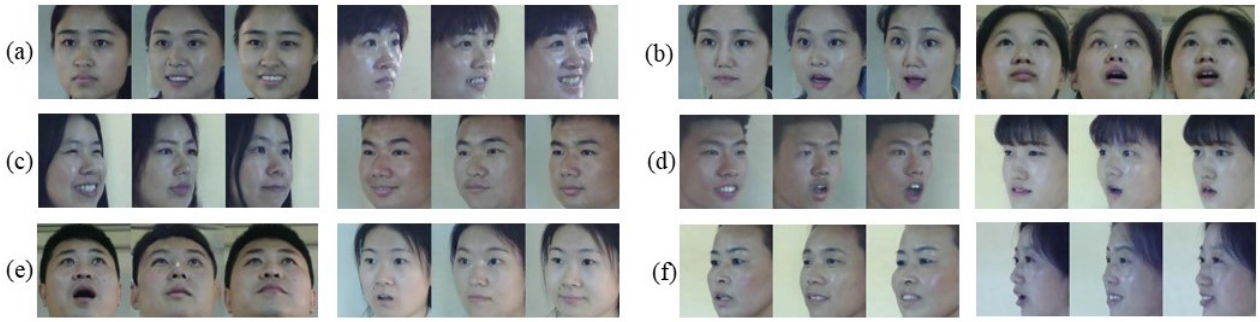


Fig. 7. Expression manipulation on M^{2}FPA. For every group of images (three images together form a group), the left, middle and right images represent the input faces, generated expressions and images with the target expressions, respectively. (a): 'Normal' to 'Happy'; (b): 'Normal' to 'Surprise'; (c): 'Happy' to 'Normal'; (d): 'Happy' to 'Surprise'; (e): 'Surprise' to 'Normal'; (f): 'Surprise' to 'Happy'.

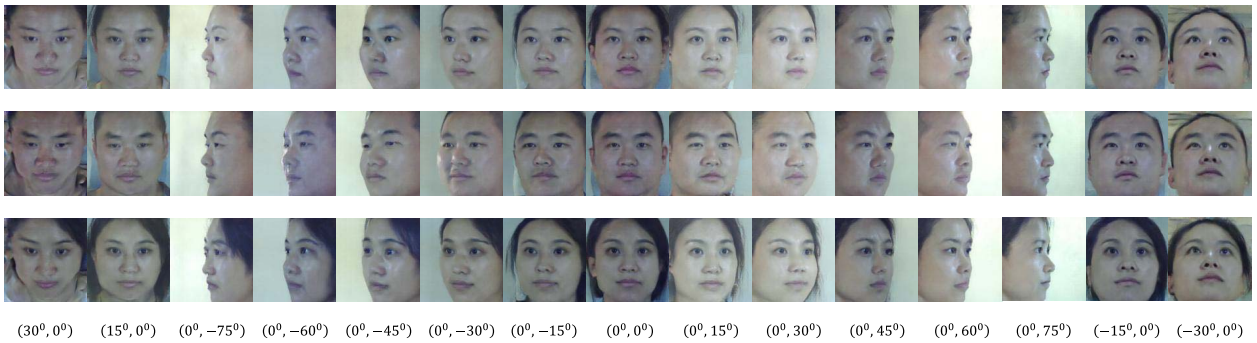


Fig. 8. Synthesis results of different target poses in the M^{2}FPA dataset, denoted as (pitch angle, yaw angle).

we introduce Z_N to supervise the embedding encoding and separation process. Note that Z_N is the feature embedding extracted by the recognition model LightCNN [108] from the normalized face $\mathbf{I}_N$, with a canonical view [51] and a neutral expression. The loss function can be termed as:

$$\mathcal{L}_{id} = ||\hat{Z}_{id} - L(\mathbf{I}_N)||_2, \tag{5}$$

where $\mathrm{L}(\cdot)$ and $\|\cdot\|_2$ denote the discriminative features extracted by LightCNN and the vector 2-norm, respectively. The same strategy is also applied while generating faces with arbitrary poses, where identity information should also be well preserved. This calls for the utilization of perceptual loss [109] over the source image $\mathbf{I}_{in}$ and the generated face image $\hat{\mathbf{I}}_{tar}$:

$$\mathcal{L}_{per} = \|L(\hat{\mathbf{I}}_{tar}) - L(\mathbf{I}_{in})\|_2, \tag{6}$$

For the second level of FDM, $\hat{Z}_1$ is further fed into the encoder E_2 to generate $\hat{Z}_2$ which denotes the complete deformation information. We separate $\hat{Z}_2$ into parts representing different attributes, i.e., $\hat{Z}_{pose}$ for pose and $\hat{Z}_{expr}$ for expression. $\hat{Z}_{pose}$ and $\hat{Z}_{expr}$ are encoded into one-hot codes $\hat{C}_{pose}$ and $\hat{C}_{expr}$ indicating different classes of poses and expressions, respectively. We impose separate constraints on the split representations to achieve disentanglement. The cross-entropy loss function is introduced to measure the differences between the predicted codes $\{\hat{C}_{expr}, \hat{C}_{pose}\}$ and the target codes $\{C_{expr}, C_{pose}\}$, which are encoded from attribute labels. The corresponding loss functions can be mathematically formulated as:

$$\mathcal{L}_{pose} = -\sum_k C^k_{pose} \log\left(\hat{C}^k_{pose}\right), \tag{7}$$

$$\mathcal{L}_{expr} = -\sum_k C^k_{expr} \log\left(\hat{C}^k_{expr}\right), \tag{8}$$

where k denotes the class indexes of the specific face attribute.

To ensure that the FDM disentangles the identities and deformations as expected, we introduce target image $\mathbf{I}_{tar}$ to supervise the generation process with a reconstruction loss function that is formulated as,

$$\mathcal{L}_{rec} = ||\hat{\mathbf{I}}_{tar} - \mathbf{I}_{tar}||_1. \tag{9}$$

The corresponding embedding triplet $< Z_{id}, C_{expr}, C_{pose} >$ is needed when the decoder generates a face image after disentanglement.

C. Overall Loss Function

1) Adversarial Loss: Our network is developed under the frame of GAN [23]. The generator G_{θ_G} generates faces with desired deformations $\hat{\mathbf{I}}_{tar}$ and the discriminator D plays a min-max game, which can be defined as:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{I}_{tar} \sim \mathbb{P}_{\text{train}}(\mathbf{I}_{tar})}\left[\log D\left(\mathbf{I}_{tar}\right)\right] + \mathbb{E}_{\mathbf{I}_{in} \sim \mathbb{P}_G(\mathbf{I}_{in})}\left[\log\left(1 - D\left(G_{\theta_G}\left(\mathbf{I}_{in}\right)\right)\right)\right]. \tag{10}$$

TABLE I

RANK-1 RECOGNITION RATES (%) ACROSS VIEWS, ILLUMINATIONS, AND SESSIONS UNDER SETTING 2 OF DATASET MULTI-PIE. TABLE A GIVES THE RESULTS OF FA-GAN COMPARED WITH OTHER STATE-OF-THE-ART METHODS. TABLE B SHOWS THE ABLATION STUDIES ON $\mathcal{L}_{per}$, THE FDM, AND THE GPM, RESPECTIVELY. OURS *w/* GPM[1] DENOTES THAT FACE IMAGES ARE CROPPED REGULARLY AS THE RIGHT ONE SHOWN IN FIG. 4. TABLE C SHOWS THE IMPACT OF VARYING PARAMETERS ON THE FINAL RESULTS

Method	±15°	±30°	±45°	±60°	±75°	±90°
A. Methods Comparison						
FIP + LDA [105]	90.7	80.7	64.1	45.9	-	-
MVP + LDA [106]	92.8	83.7	72.9	60.1	-	-
CRF [107]	95.0	88.5	79.9	61.9	-	-
DR-GAN [35]	94.0	90.1	86.2	83.2	-	-
FF-GAN [29]	94.6	92.5	89.7	85.2	77.2	61.2
TP-GAN [28]	98.7	98.1	95.4	87.8	77.4	64.6
CAPG-GAN [36]	99.8	99.6	97.3	90.6	83.1	66.1
M^{2}FPA [31]	100	99.8	99.5	96.2	88.7	75.3
RL-WGAN [99]	98.2	97.5	96.7	91.7	86.1	75.1
Ours	**100**	**100**	**99.9**	**99.9**	**94.3**	**75.9**
B. Ablation Study I						
Baseline	98.7	97.9	95.1	92.6	81.4	57.9
w/ $\mathcal{L}_{per}$	99.9	99.1	96.9	93.5	81.7	60.4
w/ $\mathcal{L}_{per}$ + FDM	100	100	99.9	98.6	91.2	73.4
Ours *w/* GPM[1]	100	100	99.9	98.9	93.6	73.8
Ours	**100**	**100**	**99.9**	**99.9**	**94.3**	**75.9**
C. Ablation Study II						
$\lambda_{pose}, \lambda_{expr} = 1$	99.9	99.8	99.5	99.6	92.4	69.9
$\lambda_{pose}, \lambda_{expr} = 0.1$	**100**	**100**	**99.9**	**99.9**	**94.3**	**75.9**
$\lambda_{pose}, \lambda_{expr} = 0.01$	100	100	99.6	99.9	95.5	72.8
$\lambda_{pose}, \lambda_{expr} = 0.001$	99.9	99.9	99.3	97.7	86.1	56.6
Ours	**100**	**100**	**99.9**	**99.9**	**94.3**	**75.9**

2) Overall Loss: The overall loss function $\mathcal{L}_{total}$ in our approach is denoted as:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{p}\mathcal{L}_{p} + \lambda_{id}\mathcal{L}_{id} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{expr}\mathcal{L}_{expr}, \quad (11)$$

where we set λ_{adv}, λ_{rec}, λ_{per}, λ_{p}, λ_{id}, λ_{pose}, and λ_{expr} to 1, 50, 1, 50, 0.01, 0.1, and 0.1, respectively.

D. Training Strategy

As discussed in the previous sections, our model is trained in a supervised manner, which means that the target face and the target face parsing map pairs, i.e., $\{\mathbf{I}_{in}, \mathbf{I}_{tar}\}$, and $\{\mathbf{I}_{in}, \mathbf{I}_{p}\}$ pairs, are needed. Note that in the training phase, $\mathbf{I}_{tar}$ alternatively takes different responsibilities, as shown in Fig. 5. The training iterations are randomly divided into two sets. For Set 1, the target images $\mathbf{I}_{tar}$ are exactly the same as the input images $\mathbf{I}_{in}$. This indicates that these iterations are responsible for reconstruction by feeding the decoder G_1 with the extracted $< \hat{Z}_{id}, \hat{C}_{expr}, \hat{C}_{pose} >$. For Set 2, $\mathbf{I}_{tar}$ maintains the same identity with randomly selected deformation attributes. Thus, we utilize $< \hat{Z}_{id}, C_{expr}, C_{pose} >$ as the input of G_1. The training strategy makes it possible for the FDM to disentangle identities with other deformation attributes, and enables the trained model to be used for generating face images with the desired deformations.

TABLE II

FACE RECOGNITION RATE (%) RESULTS FOR DATASETS IN THE WILD. THE LEFT AND RIGHT PARTS SHOW THE VERIFICATION ACCURACY ON LFW AND THE RECOGNITION RESULTS ON IJB-A, RESPECTIVELY

Method	LFW		Method	IJB-A	
	Verification			Recognition	
	ACC	AUC		Rank-1	Rank-5
TP-GAN [28]	96.13	99.42	DR-GAN [35]	85.5±1.5	94.7±1.1
LightCNN [108]	99.39	99.87	LightCNN [108]	93.0±1.0	-
CAPG-GAN [36]	99.37	99.90	FF-GAN [29]	90.2±0.6	95.4±0.5
Ours	**99.46**	**99.93**	Ours	**95.4±0.7**	**98.1±0.2**

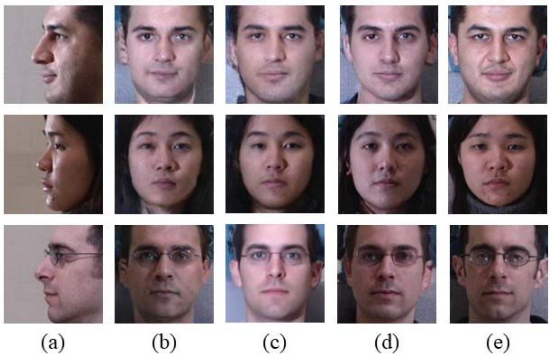


Fig. 9. Synthesis results on Multi-PIE on the pose of 90°. (a) Profiles; (b) Ours; (c) TP-GAN; (d) CAPG-GAN; (e) Canonical.

IV. EXPERIMENTS

To evaluate the performance of our proposed FA-GAN, we conduct experiments in both constrained and unconstrained environments. We consider the Multi-PIE dataset [30] and the M^{2}FPA dataset [31] which are large-scale datasets captured in constrained environments. In addition, the in-the-wild LFW [41] dataset, IJB-A dataset [19] and Megaface dataset [42] are involved in our experiments. Moreover, we utilize CASIA-WebFace and the augmented CASIA-WebFace to respectively train 8 widely used face recognition models, including ArcFace [43], CosFace [44], SphereFace [45], VGGFace [11], MobileFace [46], FaceNet [47], AdaCos [13] and CurricularFace [8]. We make comparisons with state-of-the-art methods qualitatively and quantitatively and provide in-depth analyses. Furthermore, ablation studies are conducted to evaluate the efficiency and necessity of the proposed FA-GAN, the GPM, the perceptual loss function, the different graph initialization strategies, and the values of hyperparameters.

A. Datasets

Multi-PIE [30] is developed for face-related work and includes 337 different identities, each identity is captured with 20 illumination levels and 13 poses. Following the protocol in [28], we utilize only the face images with neutral expressions under the 20 illumination levels and 13 poses in Multi-PIE. The first 200 identities are used for training,

TABLE III

RANK-1 RECOGNITION RATES (%) ACROSS VIEWS, ILLUMINATIONS, AND SESSIONS UNDER SETTING 2 OF DATASET M^{2}FPA. DEGREES ARE STRUCTURED AS ($A°$, $B°$), WHERE $A°$ DENOTES THE PITCH ANGLE AND $B°$ DENOTES THE YAW ANGLE

Method	(0°,±15°)	(0°,±30°)	(0°, ±45°)	(0°, ±60°)	(0°, ±75°)	(0°, ±90°)	(+ 15°, 0°)	(- 15°, 0°)	(+ 30°, 0°)	(- 30°, 0°)
LightCNN [108]	100	100	99.8	98.6	86.9	51.7	100	99.9	99.7	98.6
DR-GAN [35]	98.9	97.9	95.7	89.5	70.3	35.5	99.1	98.1	93.8	91.7
TP-GAN [28]	99.9	99.8	99.4	97.3	87.6	62.1	99.8	99.9	99.7	98.2
CAPG-GAN [36]	99.9	99.7	99.4	96.4	87.2	63.9	99.8	99.8	98.8	98.9
M^{2}FPA [31]	100	100	99.9	98.4	**90.6**	67.6	99.9	99.9	99.7	98.9
Ours	**100**	**100**	**99.9**	**98.7**	90.5	**68.5**	**100**	**100**	**99.7**	**99.1**

TABLE IV

VERIFICATION RESULTS OF DIFFERENT RECOGNITION MODELS ON THE LFW DATASET. FOR EACH MODEL, THE FIRST ROW ILLUSTRATES THE RESULTS OF MODELS TRAINED ON THE ORIGINAL CASIA-WEBFACE DATASET, AND THE METHOD NAMES ENDING IN "*_5w*" INDICATE THE MODELS ARE TRAINED ON THE CASIA-WEBFACE DATASET AUGMENTED WITH 50,000 IMAGES

Method	AUC (%)	EER (%)	TPR@FPR=1% (%)	TPR@FPR=0.1% (%)	TPR@FPR=0.01% (%)
ArcFace [43]	99.50	3.10	94.03	86.07	59.20
ArcFace_5w	**99.87**	**1.40**	**98.40**	**94.53**	**73.30**
SphereFace [45]	99.49	3.10	94.10	79.53	76.13
SphereFace_5w	**99.70**	**2.46**	**95.53**	**82.13**	**77.13**
CosFace [44]	99.73	1.70	97.60	91.10	81.40
CosFace_5w	**99.89**	**1.16**	**98.83**	**95.40**	**88.13**
VGGFace [11]	98.53	6.47	82.43	69.77	51.80
VGGFace_5w	**99.12**	**3.43**	**90.92**	**87.42**	**70.34**
MobileFace [46]	98.97	5.00	85.80	56.60	15.10
Mobile_5w	**99.75**	**2.00**	**96.93**	**86.80**	**80.30**
Facenet [47]	99.04	4.83	87.23	62.20	46.00
Facenet_5w	**99.46**	**3.37**	**92.03**	**78.97**	**50.73**
AdaCos [13]	99.61	**2.80**	94.27	83.70	76.63
AdaCos_5w	**99.66**	3.27	**96.17**	**89.97**	**79.73**
CurricularFace [8]	99.79	5.60	90.27	86.73	85.23
CurricularFace_5w	**99.81**	**5.49**	**90.56**	**87.13**	**86.59**

and the remaining 137 are used for testing. Hence, there is no overlap between the data for training and testing. During testing, we choose the normalized face images with canonical views and natural illumination for each identity to establish the gallery set.

M^{2}FPA [31] is a newly introduced dataset for face recognition and manipulation. It includes 229 subjects with 4 attributes and 62 poses. The protocol provided in [31] is used, which means 162 subjects are selected for training, and the remaining 67 subjects are used for testing. The gallery set is composed of the normalized face images with canonical views, neutral attributes, and above illumination. There are 105,056 and 67 images in the probe and gallery sets, respectively.

Different from Multi-PIE and M^{2}FPA, the datasets that we discuss in the following section include data from real-world scenarios. The images are crawled from the internet rather than obtained under predefined constrained environments. Thus, there are more variations in pose, expression, resolution, imaging device, and other factors, which make the datasets more challenging.

Specifically, CASIA-WebFace is a large-scale public dataset for training face recognition models collected by [20] in 2014. It contains 10,575 identities with 494,414 images in total. Because of its diversity in images and popularity in face recognition tasks, we utilize it as the training set for all our recognition models. LFW [41] is one of the earliest datasets collected in unconstrained environments, including 5,749 subjects and 13,233 images. IJB-A [19] contains 500 subjects with 5,712 images and 20,412 video frames. IJB-A is more challenging than LFW since various extreme poses are contained in the dataset. MegaFace [42] contains 690,572 identities and 1,027,060 images collected from Flickr Creative Commons. It is the first face recognition algorithm test standard at the million-scale level. All these datasets are authoritative and widely used evaluation indicators of face recognition performance. Thus, we also use them for evaluation and further comparisons.

B. Reproducibility

Before we provide the experimental details, all the images from the abovementioned datasets need to be addresseed in the same way. First, we detect their landmarks using three-level cascaded convolutional networks [21] and align them. Second, all the images are resized to 128×128 resolution. Note that we train the FA-GAN in a supervised manner. For each input image, the target face image and the corresponding face parsing map pairs, i.e., $\{\mathbf{I}_{in}, \mathbf{I}_{tar}\}$, and $\{\mathbf{I}_{in}, \mathbf{I}_{p}\}$, are needed.

TABLE V

VERIFICATION RESULTS OF DIFFERENT RECOGNITION MODELS ON THE IJB-A DATASET. FOR EACH MODEL, THE FIRST ROW ILLUSTRATES THE RESULTS OF MODELS TRAINED ON THE ORIGINAL CASIA-WEBFACE DATASET, AND THE METHOD NAMES ENDING IN "_5w" INDICATE THE MODELS ARE TRAINED ON THE CASIA-WEBFACE DATASET AUGMENTED WITH 50,000 IMAGES

Method	AUC (%)	EER (%)	TPR@FPR=1% (%)	TPR@FPR=0.1% (%)
SphereFace [45]	95.13±0.59	10.96±0.66	**73.43±1.54**	**53.73±3.66**
SphereFace_5w	**97.24±0.29**	**8.43±0.59**	73.31±2.25	47.56±3.09
CosFace [44]	97.74±0.25	7.17±0.46	76.01±3.05	44.47±6.51
CosFace_5w	**98.30±0.24**	**5.80±0.60**	**86.82±1.43**	**71.48±5.67**
VGGFace [11]	87.76±0.70	18.34±0.48	52.00±7.88	27.52±4.40
VGGFace_5w	**93.47±0.44**	**10.21±0.40**	**58.90±3.65**	**34.53±2.21**
MobileFace [46]	92.37±0.57	15.02±0.55	51.00±3.08	28.35±4.46
MobileFace_5w	**97.21±0.31**	**8.22±0.45**	**76.28±3.48**	**43.50±10.96**
Facenet [47]	93.75±0.57	11.62±0.84	72.53±2.25	55.58±2.68
Facenet_5w	**97.28±0.36**	**8.23±0.72**	**77.71±2.67**	**55.03±5.89**
AdaCos [13]	96.51±**0.40**	9.38±0.71	71.03±3.46	45.21±8.48
AdaCos_5w	**97.03**±0.54	**9.55±0.65**	**75.14±3.10**	**56.59±4.00**
CurricularFace [8]	97.39±0.60	17.25±1.00	73.63±2.34	60.11±1.70
CurricularFace_5w	**97.67±0.77**	**17.49±0.62**	**75.18±1.79**	**62.80±1.80**

TABLE VI

RECOGNITION RESULTS OF DIFFERENT RECOGNITION MODELS ON THE IJB-A DATASET. FOR EACH MODEL, THE FIRST ROW ILLUSTRATES THE RESULTS OF MODELS TRAINED ON THE ORIGINAL CASIA-WEBFACE DATASET, AND THE METHOD NAMES ENDING IN "_5w" INDICATE THE MODELS ARE TRAINED ON THE CASIA-WEBFACE DATASET AUGMENTED WITH 50,000 IMAGES

Method	Rank-1 (%)	Rank-2 (%)	Rank-3 (%)	Rank-4 (%)	Rank-5 (%)
SphereFace [45]	79.27±1.28	84.29±1.14	86.86±0.96	88.35±1.03	89.41±1.07
SphereFace_5w	**82.11±1.91**	**87.38±1.65**	**89.81±1.37**	**91.35±1.19**	**92.40±1.12**
CosFace [44]	89.46±0.85	91.97±0.66	93.14±0.59	93.71±0.72	94.22±0.76
CosFace_5w	**91.41±0.81**	**93.64±0.73**	**94.77±0.79**	**95.29±0.80**	**95.79±0.73**
VGGFace [11]	73.61±1.91	80.64±1.42	84.2±1.65	86.07±1.37	87.6±1.48
VGGFace_5w	**82.25±1.60**	**90.98±1.11**	**92.01±0.89**	**94.03±0.91**	**94.49±1.07**
MobileFace [46]	73.79±1.49	81.58±1.64	85.37±1.56	87.87±1.41	89.56±1.52
Mobile_5w	**84.29±1.32**	**87.39±1.17**	**89.07±1.12**	**90.28±1.12**	**91.00±1.08**
Facenet [47]	83.72±2.01	88.24±1.42	90.22±1.17	91.39±1.25	92.17±1.18
Facenet_5w	**85.71±1.34**	**89.91±0.96**	**92.91±0.91**	**93.35±0.86**	**94.14±0.74**
AdaCos [13]	**82.94±1.49**	87.07±**1.36**	89.20±**1.15**	90.71±1.10	91.68±1.15
AdaCos_5w	82.72±1.38	**89.67**±1.51	**90.72**±1.46	**92.74±1.09**	**93.17±1.06**
CurricularFace [8]	85.81±**1.05**	90.55±1.07	92.97±1.03	94.44±**1.03**	95.57±**1.18**
CurricularFace_5w	**86.41**±1.39	**90.80±1.04**	**93.19±0.94**	**94.68**±1.20	**95.71**±1.26

1) FA-GAN: The subnetworks architectures, including three encoders E_1, E_2, E_3 and two decoders G_1, G_2, of the proposed FA-GAN are shown in Fig. 6. The preprocessed $\mathbf{I}_{in}$ is fed into the GPM and FDM simultaneously. At the GPM branch, the cropped image regions are passed into different fully connected (FC) layers to produce one-dimensional vectors x_i as the outputs. Here, x_i is used to initialize the graph $G = \{V, E\}$ with the nodes $v_i \in V$ and the randomly predefined adjacency matrices as corresponding edges $e_i \in E$. After processing by two GCNs, we obtain the learned embedding features $\mathbf{G}''$. It is input to G_2 to generate $\hat{\mathbf{I}}_p$. Furthermore, for the other FDM branch, $\mathbf{I}_{in}$ is encoded by E_1 and E_2 to get disentangled embeddings $\hat{Z}_{id}$, $\hat{Z}_{pose}$ and $\hat{Z}_{expr}$. In addition, $\hat{\mathbf{I}}_p$ is introduced here as one of the inputs before the last BatchNorm layer of E_1 after processing by E_3.

We implement our networks using PyTorch. To evaluate the Multi-PIE and M^{2}FPA testing sets, we use the corresponding training sets to train two separate models. Our model is trained for 10 epochs with one NVIDIA Tesla V100S for approximately 8 hours, occupying 32G GPU. The computing demand is 18.5 GFLOPs in one forward modeling. Following [31], the Adam [110] optimizer is employed during training with a learning rate of 2e-4 for Multi-PIE and 1e-4 for M^{2}FPA. The batch size is set to 48.

2) Recognition Models: To make a fair evaluation, we train the evolved recognition models in the same way. ArcFace [43], CosFace [44], and SphereFace [45] share the Backbone network ResNet [111] with the cross-entropy loss. Following [43], the batch size is set to 512 for images in $128\times128\times3$. We set the initial learning rate to 0.1, which it is divided by 10 every 35 epochs. The weight decay and momentum are

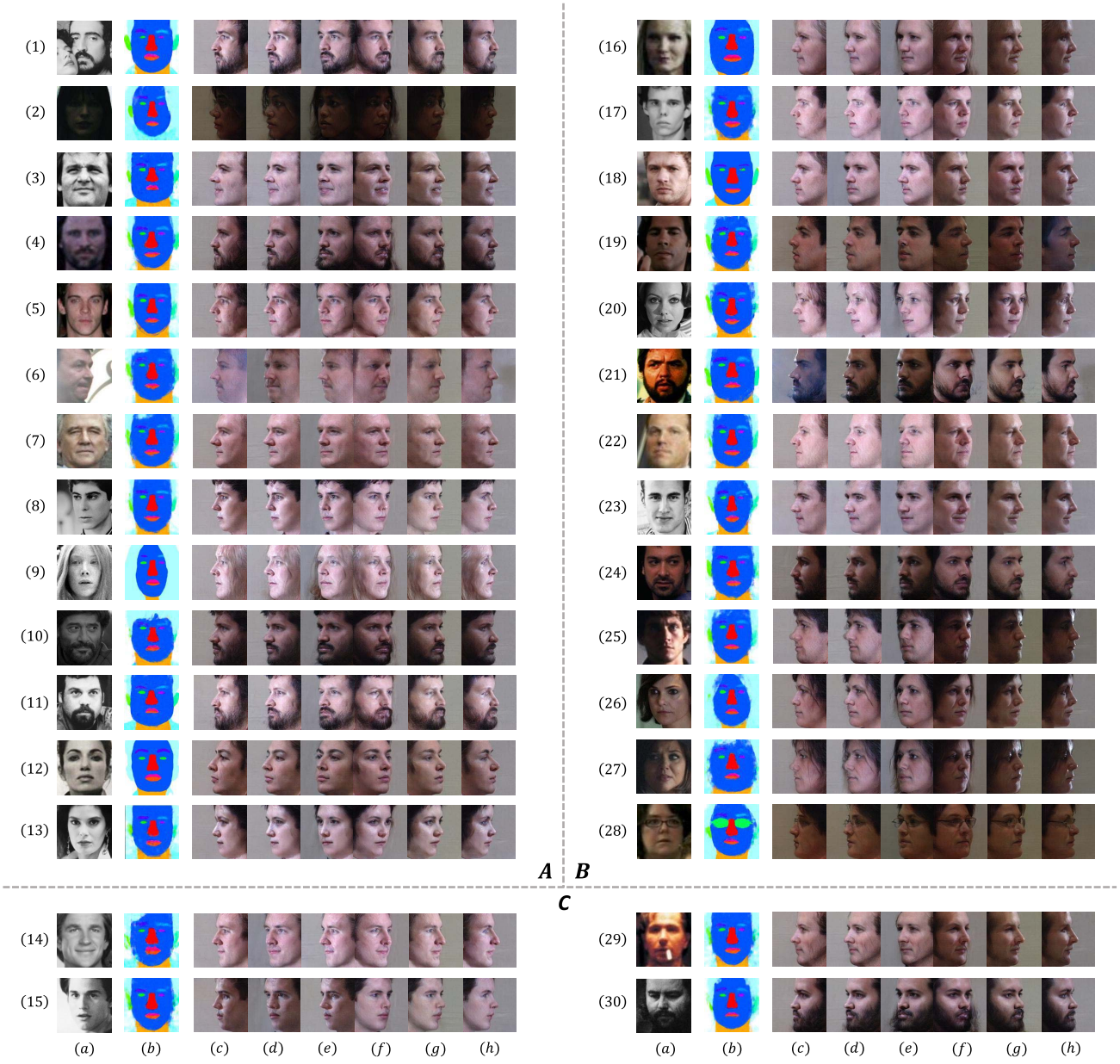


Fig. 10. Synthesis results for augmenting CASIA-WebFace. Part A and Part B give the interclass results, and Part C shows the intraclass results, where (14), (15) belong to the same person and (29), (30) belong to the same person. Moreover, Part B concentrates on the input images with uneven lighting conditions. Columns (a) and (b) are input images and predicted parsing maps respectively. Columns (c), (d), (e), (f), (g), and (h) depict the synthesized profiles with angles of $-75°$, $-60°$, $-45°$, $+45°$, $+60°$, and $+75°$, respectively.

initialized as 5e-4 and 0.9, respectively. We train the models until convergence for at most 125 epochs with one NVIDIA Tesla V100S. All of the models are trained on the original CASIA-WebFace and augmented CASIA-WebFace datasets to evaluate the performance of our FA-GAN.

C. Quantitative Analyses

Evaluating the face recognition accuracy of *recognition via generation* is a common quantitative metric to assess the effectiveness of models. In this section, we conduct extensive recognition experiments in both constrained and unconstrained environments to evaluate the identity-preserving ability of our network.

1) FA-GAN: As shown in Table I, the rank-1 recognition rate of our proposed FA-GAN on the Multi-PIE dataset is compared with those of its competitors, including FIP+LDA [105], MVP+LDA [106], CRF [107], DR-GAN [35], FF-GAN [29], TP-GAN [28], CAPG-GAN [36], M^{2}FPA [31] and RL-WGAN [99]. The results prove that our model is comparable with all the aforementioned approaches. Table III reports the detailed evaluation results on the M^{2}FPA database, where the performance of our FA-GAN is also satisfactory compared with those of DR-GAN [35], TP-GAN [28], CAPG-GAN [36] and M^{2}FPA [31].

Experiments on Multi-PIE and M^{2}FPA show that the proposed FA-GAN works well under constrained environments. Moreover, as shown in Table II, the results on the LFW and

TABLE VII

RANK-1 RECOGNITION RATES (%) WITH 1M DISTRACTORS OF DIFFERENT RECOGNITION MODELS ON MEGAFACE CHALLENGE 1 [42] USING FACESCRUB AS THE PROBE SET. THE "SMALL" PROTOCOL INDICATES THAT THE MODEL IS TRAINED WITH FEWER THAN 0.5M IMAGES. FOR EACH MODEL, THE FIRST ROW ILLUSTRATES THE RESULTS OF MODELS TRAINED ON THE ORIGINAL CASIA-WEBFACE DATASET, AND THE METHOD NAMES ENDING IN "_5w" INDICATE THE MODELS ARE TRAINED ON THE CASIA-WEBFACE DATASET AUGMENTED WITH 50,000 IMAGES

Method	Protocol	Rank-1 (%)
SphereFace [45]	Small	72.71
SphereFace_5w	Small	**73.64**
CosFace [44]	Small	77.11
CosFace_5w	Small	**84.12**
MobileFace [46]	Small	57.99
Mobile_5w	Small	**73.12**
Facenet [47]	Small	70.49
Facenet_5w	Small	**71.83**
CurricularFace [8]	Small	77.47
CurricularFace_5w	Small	**80.28**
ArcFace [43]	Small	77.50
ArcFace_5w	Small	**80.59**

IJB-A datasets further demonstrate the effectiveness of our proposed method in unconstrained environments. It can be observed that we obtain comparable results on these datasets in terms of face recognition and verification. Despite the fact that our model is trained under constrained environments, its identity-preserving ability can be well-adapted to real-life scenarios.

To verify the effectiveness of each component of the FA-GAN, we conduct ablation studies and report quantitative results. In Table I, we find that the network with the FDM outperforms the baseline. With the addition of the GPM to our model, the recognition task on Multi-PIE performs even better, suggesting that both the FDM and the GPM blocks are effective. Note that the results show that different graph initialization with different image cropping strategies also influence the performance of the GPM. We can observe that the perceptual loss function boosts the performance of the network. As shown in Table I. C, we also conduct experiments on varying parameters. We follow [36] to set λ_{adv}, λ_{rec}, λ_{per}, and λ_p, and apply grid search to tune other parameters.

2) Recognition Models: We train the selected recognition models with the original CASIA-WebFace and the augmented CASIA-WebFace separately, and the results are depicted in Table IV, V, VI, and VII. It is noticeable that all the recognition models acquire performance improvements in terms of recognition rates, Equal Error Rate (EER) and Area Under Curve (AUC) to some extent. Most models achieve a dramatic increase in True Positive Rate (TPR). For example, TPR@FAR $= 0.1\%$ of CosFace increases from 44.47% to 71.48%, which is a relative improvement of 60%. MobileFace performs even better in terms of TPR@FAR $= 0.1\%$, increasing from 28.35% to 43.5%, which is a relative improvement of 70%. The variances of these evaluation metrics on most recognition models also decrease to varying degrees, indicating that the models become more stable after being trained on the augmented datasets. Although the TPR of SphereFace decreases slightly, it falls within an acceptable margin.

D. Qualitative Analyses

To qualitatively analyze the performance of our proposed FA-GAN, we further provide the visualization results. Fig. 7 demonstrates the outstanding ability of our proposed network to disentangle and manipulate expressions under different poses on the M^2FPA dataset. Fig. 8 gives the synthesis results of different target poses, demonstrating the identity-preserving ability of the FA-GAN. For comparison, the results of the state-of-the-art methods on Multi-PIE can be found in Fig. 9. The generated faces used for augmenting CASIA-WebFace can be found in Fig. 10, where Parts A and B demonstrate the interclass results while Part C gives the interclass results. For each input image, we give the predicted parsing map of the corresponding normalized face and the synthesized results of 6 different poses. It can be found that our model is also robust under occlusion. For example, for input (1), the FA-GAN identifies the target person accurately with regard to the lady as background, which should be credited to GPM. For input (28), the glasses belong to deformation-irrelevant attributes and are well preserved as part of geometry information. Moreover, the network is also skilled in managing uneven lighting, as shown in Fig. 10 Part B. It can be observed that the left faces and right faces of these images are under different lighting conditions, and the generated profiles adapt to these conditions appropriately.

V. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel graph-based two-stage Face Augmentation Generative Adversarial Network to augment existing datasets for deformation-invariant face recognition. It not only disentangles the identity representations to improve the face recognition accuracy, but also utilizes the disentangled representations to manipulate face attributes. We also introduce Graph Convolutional Networks to explore high-level interrelations between different face regions to better preserve the geometric information. Extensive experiments are conducted on face recognition and face synthesis tasks to demonstrate that our proposed network acquires a good identity-preserving ability from constrained datasets. This ability is also well-adapted to real-life environments while manipulating faces with desired deformations; thus, proving the effectiveness and generalization ability of our approach.

In the future, we would like to make the framework more flexible and generalized for deformation-invariant person re-identification. Moreover, considering the trend of using less data, transferring the FA-GAN from a fully-supervised framework to a semi-supervised one would also be interesting.

ACKNOWLEDGMENT

The authors would like to thank the associate editors and reviewers for their valuable suggestions.

REFERENCES

[1] A. Dabouei, F. Taherkhani, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Boosting deep face recognition via disentangling appearance and geometry," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 320–329.

[2] H. Qin, "Asymmetric rejection loss for fairer face recognition," 2020, *arXiv:2002.03276*. [Online]. Available: http://arxiv.org/abs/2002.03276

[3] P. Terhörst, J. Niklas Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," 2020, *arXiv:2002.03592*. [Online]. Available: http://arxiv.org/abs/2002.03592

[4] Y. Huang *et al.*, "Improving face recognition from hard samples via distribution distillation loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 138–154.

[5] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 0–1.

[6] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6817–6826.

[7] D. Cao, X. Zhu, X. Huang, J. Guo, and Z. Lei, "Domain balancing: Face recognition on long-tailed domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5671–5679.

[8] Y. Huang *et al.*, "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5901–5910.

[9] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-face: Dual variational generation for heterogeneous face recognition," 2020, *arXiv:2009.09399*. [Online]. Available: http://arxiv.org/abs/2009.09399

[10] X. Yin, Y. Tai, Y. Huang, and X. Liu, "FAN: Feature adaptation network for surveillance face recognition and normalization," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 1–17.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.

[12] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5089–5097.

[13] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10823–10832.

[14] Y. Duan, J. Lu, and J. Zhou, "UniformFace: Learning deep equidistributed representation for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3415–3424.

[15] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "AdaptiveFace: Adaptive margin and sampling for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11947–11956.

[16] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometric recognition using deep learning: A survey," 2019, *arXiv:1912.00271*. [Online]. Available: http://arxiv.org/abs/1912.00271

[17] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2080–2089.

[18] X. Lu and A. K. Jain, "Deformation analysis for 3D face matching," in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, vol. 1, Jan. 2005, pp. 99–104.

[19] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[20] 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

[21] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[22] A. Bingham and D. Spradlin, *The Long Tail Expertise*. London, U.K.: Pearson, 2011.

[23] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

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

[25] A. Van den Oord *et al.*, "Conditional image generation with PixelCNN decoders," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 4790–4798.

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[27] 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.

[28] 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.

[29] 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. 3990–3999.

[30] 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.

[31] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10043–10051.

[32] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1999, pp. 187–194.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[34] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, "Unsupervised depth estimation, 3D face rotation and replacement," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9736–9746.

[35] 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, pp. 1415–1424.

[36] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[37] Q. Li, X. Zhao, R. He, and K. Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8634–8641.

[38] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 103–118.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[40] Y. Li, H. Huang, J. Cao, R. He, and T. Tan, "Disentangled representation learning of makeup portraits in the wild," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2166–2184, Sep. 2020.

[41] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. inria-00321923, 2008.

[42] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 Million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.

[43] 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.

[44] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[45] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[46] N. Chinaev, A. Chigorin, and I. Laptev, "MobileFace: 3D face reconstruction with efficient CNN regression," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 15–30.

[47] 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.

[48] 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.

[49] 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.

[50] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1047–1052.

[51] M. J. Owen, "Simple canonical views," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 839–848.

[52] 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.

[53] J. Zhao *et al.*, "Dual-agent GANs for photorealistic and identity preserving profile face synthesis," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 66–76.

[54] X. Ma, X. Zhou, H. Huang, Z. Chai, X. Wei, and R. He, "Free-form image inpainting via contrastive attention network," 2020, *arXiv:2010.15643*. [Online]. Available: http://arxiv.org/abs/2010.15643

[55] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "3D aided duet GANs for multi-view face image synthesis," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2028–2042, Aug. 2019.

[56] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Towards high fidelity face frontalization in the wild," *Int. J. Comput. Vis. (IJCV)*, vol. 128, pp. 1485–1504, Oct. 2019.

[57] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 2867–2877.

[58] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 1099–1107.

[59] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[60] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*. [Online]. Available: http://arxiv.org/abs/1809.11096

[61] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[62] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proc. CVPR*, Jun. 2011, pp. 625–632.

[63] S. Minaee, A. Abdolrashidi, and Y. Wang, "Face recognition using scattering convolutional network," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2017, pp. 1–6.

[64] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.

[65] R. He, B.-G. Hu, and X.-T. Yuan, "Robust discriminant analysis based on nonparametric maximum entropy," in *Proc. Asian Conf. Mach. Learn. (ACML)*. Berlin, Germany: Springer, 2009, pp. 120–134.

[66] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000.

[67] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2002, pp. 447–460 .

[68] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2011, pp. 44–51.

[69] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," 2014, *arXiv:1412.6583*. [Online]. Available: http://arxiv.org/abs/1412.6583

[70] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," 2012, *arXiv:1210.5474*. [Online]. Available: http://arxiv.org/abs/1210.5474

[71] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High fidelity face manipulation with extreme poses and expressions," 2019, *arXiv:1903.12003*. [Online]. Available: http://arxiv.org/abs/1903.12003

[72] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 5040–5048.

[73] C. Wang, C. Xu, and D. Tao, "Tag disentangled generative adversarial networks for object image re-rendering," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2901–2907.

[74] 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. (NeurIPS)*, 2016, pp. 2172–2180.

[75] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. SIDL-WP-1999-0120, 1999.

[76] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Comput. Surv.*, vol. 31, no. 4es, p. 5, Dec. 1999.

[77] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 3844–3852.

[78] L. Yi, H. Su, X. Guo, and L. Guibas, "SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2282–2290.

[79] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 3189–3197.

[80] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on Riemannian manifolds," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 37–45.

[81] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5115–5124.

[82] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Nov. 2019.

[83] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*. [Online]. Available: http://arxiv.org/abs/1312.6203

[84] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 399–417.

[85] P. Novák, P. Neumann, and J. Macas, "Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data," *BMC Bioinf.*, vol. 11, no. 1, p. 378, 2010.

[86] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[87] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.

[88] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019.

[89] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*. [Online]. Available: http://arxiv.org/abs/1805.09501

[90] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, and L. Franco, "Forward noise adjustment scheme for data augmentation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 728–734.

[91] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[92] A. Jurio, M. Pagola, M. Galar, C. Lopez-Molina, and D. Paternain, "A comparison study of different color spaces in clustering based image segmentation," in *Proc. Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst. (IPMU)*, 2010, pp. 532–541.

[93] T. Konno and M. Iwazume, "Icing on the cake: An easy and quick post-learnig method you can try after deep learning," 2018, *arXiv:1807.06540*. [Online]. Available: http://arxiv.org/abs/1807.06540

[94] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," 2017, *arXiv:1702.05538*. [Online]. Available: http://arxiv.org/abs/1702.05538

[95] C. Bowles *et al.*, "GAN augmentation: Augmenting training data using generative adversarial networks," 2018, *arXiv:1810.10863*. [Online]. Available: http://arxiv.org/abs/1810.10863

[96] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.

[97] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 579–596.

[98] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter, "Training deep face recognition systems with synthetic data," 2018, *arXiv:1802.05891*. [Online]. Available: http://arxiv.org/abs/1802.05891

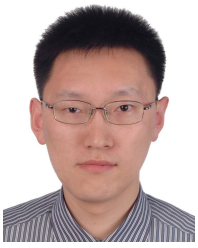
[99] C.-H. Tang, G.-S.-J. Hsu, and M. Hoon Yap, "Face recognition with disentangled facial representation learning and data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1670–1674.
[100] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017.
[101] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation," 2017, *arXiv:1708.06020*. [Online]. Available: http://arxiv.org/abs/1708.06020
[102] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[103] H.-Y. Lee *et al.*, "DRIT++: Diverse Image-to-Image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2402–2417, Nov. 2020.
[104] T. Xiao, J. Hong, and J. Ma, "ELEGANT: Exchanging latent encodings with GAN for transferring multiple face attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 168–184.
[105] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 113–120.
[106] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 217–225.
[107] 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.
[108] 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.
[109] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.
[110] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[111] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

Mandi Luo received the B.E. degree in automation engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the B.Sc. and M.Sc. degrees in electronic engineering from the Katholieke Universiteit Leuven, Leuven, Belgium, in 2017 and 2018, respectively. She is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences, Beijing, China. Her research interests include biometrics, pattern recognition, and computer vision.

Jie Cao received the B.E. degree in automation from North China Electric Power University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include biometrics, pattern recognition, computer vision, and machine learning.

Xin Ma received the B.E. degree in electronic information engineering from Jiangsu University (JSU), Jiangsu, China, in 2018. He is currently pursuing the M.S. degree in computer technology with the University of Chinese Academy of Sciences (UCAS), Beijing, China. His research interests include image super-resolution, image inpainting, and machine learning.

Xiaoyu Zhang received the B.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. His research interests include machine learning, data mining, and big data analysis. His awards and honors include the Silver Prize of Microsoft Cup IEEE China Student Paper Contest in 2009, the Second Prize of the Wu Wen-Jun AI Science and Technology Innovation Award in 2016, the CCCV Best Paper Nominate Award in 2017, and the Third Prize of BAST Beijing Excellent S&T Paper Award in 2018.



Ran He (Senior Member, IEEE) received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, Dalian, China, in 2001 and 2004, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2009. Since September 2010, he has been joined NLPR, where he is currently a Full Professor. His research interests include information theoretic learning, pattern recognition, and computer vision. He serves as an Associate Editor for the *Neurocomputing* (Elsevier), and also serves on the program committee for several conferences.