Geometry Guided Pose-Invariant Facial Expression Recognition

Feifei Zhang, Tianzhu Zhang, *Member, IEEE*, Qirong Mao, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— Driven by recent advances in human-centered computing, Facial Expression Recognition (FER) has attracted significant attention in many applications. However, most conventional approaches either perform face frontalization on a non-frontal facial image or learn separate classifier for each pose. Different from existing methods, this paper proposes an end-to-end deep learning model that allows to simultaneous facial image synthesis and pose-invariant facial expression recognition by exploiting shape geometry of the face image. The proposed model is based on generative adversarial network (GAN) and enjoys several merits. First, given an input face and a target pose and expression designated by a set of facial landmarks, an identity-preserving face can be generated through guiding by the target pose and expression. Second, the identity representation is explicitly disentangled from both expression and pose variations through the shape geometry delivered by facial landmarks. Third, our model can automatically generate face images with different expressions and poses in a continuous way to enlarge and enrich the training set for the FER task. Our approach is demonstrated to perform well when compared with state-of-the-art algorithms on both controlled and in-the-wild benchmark datasets including Multi-PIE, BU-3DFE, and SFEW. The code is included in the supplementary material.**

***Index Terms*— Facial expression recognition, facial image synthesis, generative adversarial network, facial landmarks.**

Manuscript received October 30, 2018; revised July 13, 2019; accepted January 23, 2020. Date of publication February 12, 2020; date of current version February 21, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002804, in part by the National Natural Science Foundation of China (NSFC) under Grant 61720106006, Grant 61721004, Grant 61832002, Grant 61532009, Grant U1705262, Grant U1836220, Grant 61702511, Grant 61672267, and Grant 61751211, in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039, and in part by the Research Program of National Laboratory of Pattern Recognition under Grant Z-2018007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Senem Velipasalar. *(Corresponding author: Changsheng Xu.)*

Feifei Zhang was with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212000, China. She is now with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: susanzhang@ujs.edu.cn).

Tianzhu Zhang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: tzzhang10@gmail.com).

Qirong Mao is with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212000, China (e-mail: mao_qr@ujs.edu.cn).

Changsheng Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: csxu@nlpr.ia.ac.cn).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TIP.2020.2972114

I. Introduction

Facial behavior is one of the most important channels for emotional communication between humans. Automatic facial expression recognition (FER) has attracted increasing attention in recent years because of its wide potential applications in psychology, medicine, security, digital entertainment, and driven monitoring, to name but a few [1]–[5]. Due to subtle facial appearance changes and significant subject-dependent variations, the FER is a rather challenging task. Despite of significant progress in recent years, it remains enduring challenges for developing robust algorithms to recognize facial expression in scenarios with challenging factors such as the high nonlinearity of facial expression changes, large pose variations, variances of individuals, and insufficient training data.

Facial expression recognition aims to analyze and classify a given facial image into several emotion types, such as the Ekman's six universal expression categories (i.e., anger, disgust, fear, happiness, sadness, and surprise) [6]. As a step toward this goal, numerous algorithms have been proposed for the FER task [7]–[9], which can give stunning results on frontal or nearly frontal view facial images. However, as shown in Figure 1, the facial images are always taken from multiple views and exhibit spontaneous, which makes non-frontal or in the wild FER more challenging and thus largely unexplored. In contrast to the frontal FER, expression recognition from non-frontal facial images is challenging because it needs to deal with the issues of face occlusions, accurate non-frontal face alignment, and accurate non-frontal facial points location as shown in Figure 1. As a result, only a small part of algorithms have been proposed to address this challenging issue [10]–[12]. Especially, some of them [13], [14] have trained multiple models for each specific pose and thus need parameter-tuned separately for each model, which is time-consuming. Different from existing methods, we focus on the pose-invariant FER in an end-to-end manner, which is to perform FER by identifying or authorizing individual's expressions with facial images captured under arbitrary poses, and the separate training and parameter tuning for each pose is not required. Therefore, it is more challenging and more applicable in real scenarios.

This new task has new challenges. As shown in Figure 1, facial expression is the result of the combined and coordinated action of facial muscles, which have large variations under different poses. Clearly, it is hard to perform decoupling of the non-rigid facial changes due to the expression and rigid facial

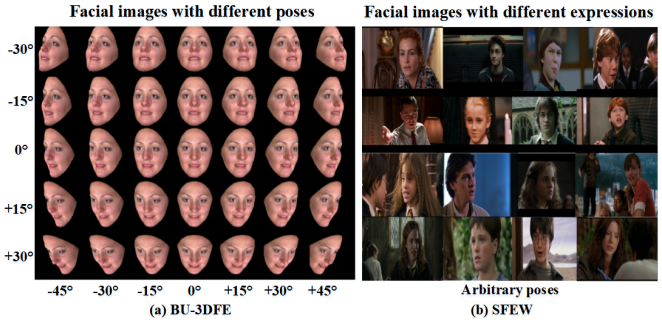


Fig. 1. Facial expression recognition is a challenging task due to the varied head poses, unconstrained facial expressions, and insufficient training examples. We aim to disentangle the expression, pose and identity from the facial image to generate more labeled training data and train a pose-invariant facial expression recognition model.

changes due to the head-pose, as they are non-linearly coupled in 2D images [15]. In details, the shape of facial texture is warped nonlinearly along with the pose change, which causes serious confusion with the inter-personal texture difference. Besides, head rotation results in self-occlusion, which causes information loss for facial expression recognition. Thus it is necessary to do head-pose analysis and facial expression jointly. Nonetheless, because of the large appearance variation of facial expressions under different poses, it is still challenging to develop an automated system that can accurately decouple these two sources of variation. Existing methods that address the above issues can be divided into three categories. **(1)**, extract pose-invariant features as facial expression representation and employ a single conventional classifier for recognition. Here, traditional methods often make use of robust local descriptors such as local binary pattern (LBP) [16], histograms of oriented gradients (HOG) [16], and scaled-invariant feature transform (SIFT) [17] to account for local distortions and then adopt classification techniques, e.g., Support Vector Machine (SVM), to achieve pose-invariant facial expression recognition. In [17], the region covariance matrix is extracted by computing the covariance of SIFT vectors which are extracted from each facial image. However, due to the limited representation power and the tradeoff between invariance and discriminability, these approaches may not deal with the challenge of nonlinear facial texture warping caused by large pose variation well. **(2)**, perform pose normalization to learn a mapping between the frontal and non-frontal facial images or to recover a frontal view image from a large pose face image, and then use the frontalized facial features or recovered face images for pose-invariant facial expression recognition. In [18], 39 landmark points are located from each non-frontal head pose, and a Gaussian process regression model is adopted to exploit pairwise correlations among different poses in order to learn robust mappings from non-frontal pose to frontal one. These methods are good at normalizing small pose faces. However, their performance decreases under large face poses due to their strong dependence on the accuracy of facial landmark detection. In [19], Zhang et al. propose to utilize 3D geometrical transformations to render a frontal view by aligning the 2D facial expression image with a 3D model. Although their results are encouraging, the pose normalization is tackled separately from the FER task, rather than an end-to-end manner, which is inconvenient and complicated. **(3)**, learn multiple classifiers for each specific pose. Most methods usually utilize view-specific classifier (e.g., view-specific SVMs) for each view [13], [14]. However, it is time-consuming with the increase of pose number, which makes these methods suboptimal for the pose-invariant facial expression recognition.

TABLE I

THE DETAILS OF EXISTING BENCHMARKS FOR POSE-INVARIANT FER INCLUDING THE NUMBER OF POSE, EXPRESSION, AND TRAINING SAMPLES

Dataset	Pose	Expression	Training Samples
SFEW	-	7	700
Multi-PIE	5	6	7,655
BU-3DFE	35	6	21,000

Recently, several researchers have successfully leveraged deep networks in a wide range of visual tasks and got promising results, such as image classification [20], object detection [21], segmentation [22], and pose estimation [23]. Inspired by the success of deep networks, an intuitive idea is to learn semantic features for the FER via deep learning. However, a prerequisite of deep model training is the availability of large-scale labeled training data to make the model robust to the variations in natural images, whereas current facial expression datasets typically contain a very limited number of labelled samples [24]. Thus, the first step in creating such a successful facial expression classification system is gathering sufficient annotated data where each image is labeled with the correct category. Nevertheless, for the pose-invariant FER, the publicly available datasets typically contain a very limited number of labeled samples. As shown in Table I, for example, the Static Facial Expressions in the wild (SFEW) dataset [25] contains only 700 images (including both training and testing) while the Multi-PIE [26] has 7,655 images (5 poses and 6 expressions). Although the 3D facial expression dataset (BU-3DFE) [27] is larger than them, it only has 21,000 images (for 35 poses). Furthermore, collecting and manually annotating such data is laborious and error prone, especially for large-scale datasets.

An avenue for overcoming the lack of labeled training data is to adopt deep networks pre-trained on ImageNet [28] and do fine tuning to further improve the feature representation power. As a consequence, the networks are trained separately from the FER, and the extracted features hardly benefit from the end-to-end training. In recent years, computational models based on end-to-end learnable convolutional networks have made significant improvements for visual recognition [29] than the methods that trained individual components separately. The reason is that in the end-to-end structure, the free parameters in all components can co-adapt and cooperate to achieve a single objective by the task-specific loss. Another solution is to generate training data automatically. Recently, GAN models have been particularly popular because of their principle ability to generate sharp images through adversarial training. GAN-based approaches have been successfully used in a

wide range of applications, such as generating house numbers [30], flowers [31], birds [32], and so on. GANs have also been widely applied in face-related tasks, such as face pose manipulation [33], face aging [34], and facial expression synthesis [35]. This inspires us to resort to the GAN to enlarge and enrich the training set. Despite many promising developments [32], [35]–[37], image synthesis remains the main objective of GAN, which cannot be straightforwardly applied to our FER task. Besides, existing works related to GAN-based facial expression synthesis mainly focus on generating facial expressions of annotated categories in facial expression datasets. However, from the perspective of psychology, facial expressions are the result of the combinations of muscle movements underneath the skin of the face that cannot be categorized in a discrete and low number of classes.

To address the above issues, on one hand, this paper proposes a geometry guided GAN-based structure to generate labeled facial images with arbitrary expressions and poses conditioned on a set of facial landmarks. On the other hand, we embed a classifier into the network to facilitate the image synthesis and conduct facial expression recognition. Our method consists of a facial geometry embedding network, an image generator network, two image discriminator networks, and an expression classifier. In particular, the geometry embedding network is used to map facial landmarks onto a semantic manifold. Besides, to disentangle the attributes (expression, pose) from the identity representation, we construct a generator G with an encoder-decoder structure, which serves as a facial image changer. The input to the encoder G_{enc} is a face image of any expression and pose, and it learns a mapping from the input facial image to a feature representation. The representation is then concatenated with the geometry information to feed into G_{dec}. The output of the decoder G_{dec} is a synthetic facial image with a target expression and pose, and the learnt identity representation bridges G_{enc} and G_{dec}. Furthermore, we introduce two discriminators (D_{att} and D_i) into the generative adversarial network, which are designed for attributes and identity features, respectively. The D_{att} is used to disentangle the pose, expression and identity from a facial image in a latent space to change the attributes (expression, pose) but retain the identity. To smooth the pose and expression transformation, the D_i is adopted to control the distribution of identity features. With an additional classifier C_{exp}, it can strive for the generated facial image to have the same expression as the input real facial image, which mainly has two effects. First, it acts as a facial expression classification model that outputs the expression label for the facial images. Second, the learnt representation is more generative to synthesize an identity-preserving facial image but with different expressions and poses, and the generated facial images can facilitate the FER in turn.

A preliminary version of this work was published in [38]. We extend it in numerous ways: 1) Instead of using the discrete pose and expression codes to synthesize the new facial image, we leverage the geometry information designed by a set of facial landmarks to represent the pose and expression, which makes the model more flexible and can generate face images with different expressions and poses in a continuous way. 2) We add a facial geometry embedding network to extract the geometry vector. 3) We conduct all experiments using the new model, and obtain better experimental results. The main contributions of our work lie in three folds:

1). We propose an end-to-end formulation to achieve facial image synthesis and pose-invariant facial expression recognition jointly. The proposed model demonstrates the possibility to synthesize photorealistic and identity-preserving faces with arbitrary expressions and poses and achieves state-of-the-art facial expression recognition performance on Multi-PIE [25], BU-3DFE [27], and SFEW [25] datasets.

2). The identity representation learning is explicitly disentangled from both expression and pose variations through the geometry information in G and D. As a result, the proposed model can automatically generate labeled facial images with an arbitrary expression under an arbitrary pose, which could explicitly facilitate our final expression recognition task.

3). The facial landmarks are embedded as a controllable signal in geometry guided generator to synthesize facial image. Compared to the previous method directly based on discrete pose and expression categories or 3D information, the facial landmarks provides a flexible and efficient way for both learning and inference. Experimental results demonstrate that our proposed method can be applied in the smooth transition between different facial expressions through geometry interpolation.

II. Related Work

Facial expression recognition has been studied extensively over the past decades. A comprehensive review of the FER methods is beyond the scope of the paper, and surveys of this field can be found in [39]. In this section, we discuss the methods closely related to this work in terms of the generative adversarial networks (GANs), Conditional GANs, and facial expression recognition.

A. Generative Adversarial Networks

Since the Generative Adversarial Network (GAN) is proposed by Goodfellow *et al.* [40], researchers have studied it vigorously. The GAN is a powerful class of generative model based on a minimax two-player game theory. The optimization of a typical GAN, as shown in Figure 3 (a), consists in simultaneously training a generator network G to produce realistic fake samples and a discriminator network D trained to distinguish between real and fake data. This idea is embedded by the so-called *adversarial loss*, and through this game, the generator and discriminator can both improve themselves. Concretely, D and G play the game with a value function $V(D, G)$:

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

where $p_d(x)$ denotes the real distribution of images x, and $p_z(z)$ is the prior on the input noise variables z. The two

parts, G and D, are trained alternatively. When the adversarial process reaches the Nash equilibrium, the minimax game attains its global optimum $p(G(z)) \sim p_d(x)$. Upon convergence, D can reject images that look fake, and G can produce high-quality images, which can fool D. A variety of GANs have been proposed for image translation [41], face generation [3], super-resolution imaging [42], indoor scene modeling [43], and human poses editing [44]. Specifically, CycleGAN [41] introduces a cycle consistency loss to learn a mapping from a source domain to a target domain without paired training examples. Yang *et al.* [3] propose a De-expression Residue Learning (DeRL) model to generate the corresponding neutral face image for any input face image. Although the GAN models have been proven to produce realistic images with a high level of detail, one of the biggest issues of GAN is that the training process is unstable, and the generated images are often noisy and incomprehensible.

B. Conditional GAN (cGAN)

The cGAN [45] is an extension of the GAN, which is developed for the conditional generating. Specifically, the cGAN extends GAN by feeding the labels y to both G and D to generate images conditioned on the labels, either class labels, modality information, or even partial data for inpainting. The objective function of the conditional GAN can be rewritten as:

$$\min_G \max_D V(D,G) = E_{x,y\sim p_d(x,y)}[\log D(x,y)] + E_{z\sim p_z(z),y\sim p_y(y)}[\log(1-D(G(z,y),y))]. \quad (2)$$

where $p_y(y)$ denotes the distribution of the labels. Prior studies have explored combining several conditions, such as text descriptions [37], [40] and class information [38]. The most relevant methods to our work are about face generation, such as face attribute editing [46], face frontalization [47], [48], facial image supervision [49], and face completion [50]. For example, Zhou and Shi [51] propose a conditional difference adversarial autoencoder (CDAAE) to generate faces conditioned on emotion classes or AU labels. Choi *et al.* [35] propose the StarGAN to perform image-to-image translations for multiple domains using only a single model, which can also be applied in facial expression synthesis. Previous studies mainly focus on generating facial images conditioned on discrete emotion classes. However, human emotion is expressed in a continuous way, thus discrete labels are not sufficient to describe detailed characteristics of facial expression. Besides, our method can explicitly disentangle the identity representation learning from both expression and pose variations by using the facial landmarks.

C. Facial Expression Recognition

Extensive efforts have been devoted to recognizing facial expressions [2], [12], [52]–[55]. Most of existing methods on the FER study the expressions of six basic emotions including happiness, sadness, surprise, fear, anger and disgust because of their marked reference representation in our affective lives and the availability of the relevant training and test data [56]. Generally, the learning system mainly includes two stages, i.e., feature extraction and expression recognition. In the first stage, features are extracted from facial images to characterize facial appearance/geometry changes caused by activation of a target expression. According to whether the features are extracted by manually designed descriptors or by deep learning methods, they can be grouped into engineered features [10], [57], [58] and learning-based features [8], [9], [59], [60]. For the engineered features, it can be further divided into texture-based local features, geometry-based global features, and hybrid features. The texture-based features mainly include SIFT [57], HOG [21], Histograms of LBP [16], Haar features [61], and Gabor wavelet coefficients [62]. The geometry-based global features are mainly based on the landmark points around eyes, mouth, and noses [18], [63]. The hybrid features usually refer to the features by combining two or more of the engineered features [10]. The learning-based features are based on deep neutral networks [9], [64]. Not surprisingly, almost all of them use some form of unsupervised pre-training/learning to initialize their models. It is mainly because the scarcity of labeled data prevent the authors from training a completely supervised model due to the overfitting problem. The most direct and effective solution to this problem is manually labeling more data. However, it may be infeasible for the FER with arbitrary poses. After feature extraction, in the next stage (expression classification), the extracted features are fed into a supervised classifier, e.g., Support Vector Machines (SVMs) [13], softmax [60], or logistic regression [64], to train a facial expression recognizer for a target expression. Different from existing methods, we use a variation of GAN to automatically generate facial images with different expressions and poses. Furthermore, our classifier is trained with the GAN in an end-to-end manner.

III. Proposed Method

In this section, we first introduce the architectures of the proposed network for simultaneous facial image synthesis and pose-invariant FER. Then, we show the learning process in details and discuss the difference with existing models.

A. Geometry Guided Pose-Invariant FER

We propose an end-to-end learning model by exploiting different facial landmarks for simultaneous facial image synthesis and pose-invariant facial expression recognition. The architecture of our model is shown in Figure 2. Before passing an image into our model, the faces and facial landmarks are obtained through dlib [65] and the number of the facial landmarks is 68 in total including landmark points of two eyebrows, two eyes, the nose, the lips, and the jaw. To synthesize a facial image of any expression and pose, our framework requires two input information, i.e., one input facial image x, and the target facial landmarks g^t. Our network synthesizes a new face image x' of the subject with the target facial landmarks, which could facilitate our final FER task.

As shown in Figure 2, our model is comprised of five basic subnetworks, i.e., a facial geometry embedding network E, a generator G, two discriminators D_i and D_{att}, and a

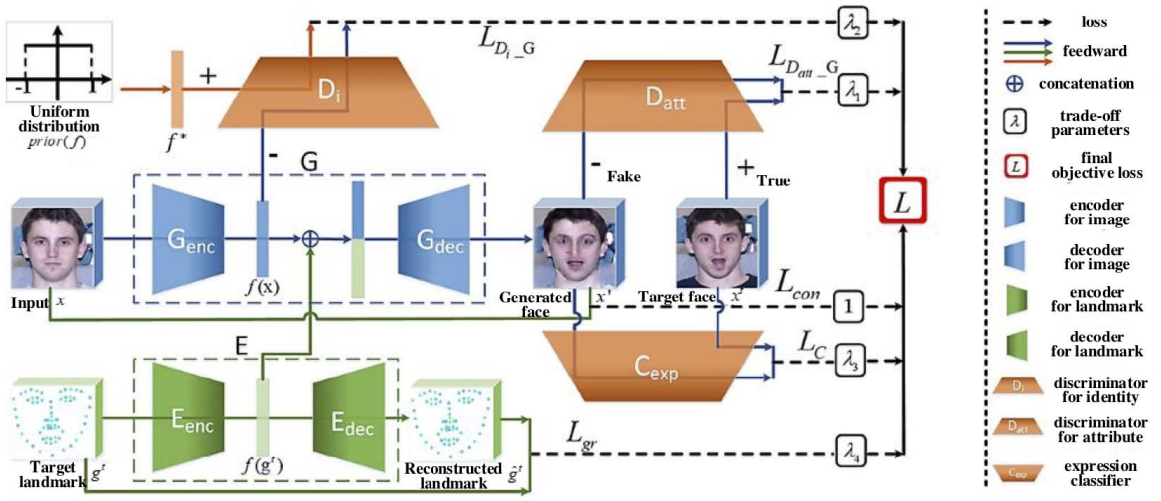


Fig. 2. The overall architecture of the proposed model, which incorporates a generator G, two discriminators D_{att} and D_i, a classifier C_{exp}, and a geometry embedding network E. Conditioned on a set of facial landmarks, the proposed model can generate facial images with different expressions under arbitrary poses to enlarge and enrich the training set for the FER task. L_* means the loss of the corresponding sub-module. $+$ and $-$ mean the input data is regarded as true or fake in the discriminator.

classifier C_{exp}. The function of the *network* E is to extract the geometry vector $f(g^t)$. The *generator* G is designed as an encoder-decoder structure to generate the new facial image x'. Specifically, the G_{enc} is used to learn a mapping from the input image to the identity feature representation $f(x)$. The representation is then concatenated with the target geometry embedding $f(g^t)$ to feed to G_{dec} for face changing. The *network* D_{att} is adopted to distinguish the real from generated facial image, which is helpful for disentangling the pose, expression and identity from a facial image in a latent space to change the attributes (pose and expression) but retain the identity. The *network* D_i is adopted to control the distribution of identity features $f(x)$, and smooth the pose and expression transformation, which could improve the quality of the generated images and has been demonstrated effectively in [66]. The *classifier* C_{exp} is used to classify the expression by measuring the posterior probability $P(c|x)$, where c is the expression label of x. We adopt a deep modeling approach for the classifier, which guarantees that, at each layer, the features become increasingly invariant to nuisance factors while maintaining discriminative information with respect to the task of facial expression recognition.

B. Training Losses

Given a facial image x following distribution p_i with label y^e and a target facial image x^t and facial landmark g^t corresponding to label y^t, where y^e and y^t represents the label for expression following distribution p_y, the objectives of our learning problem are threefold: (1) Disentangle the identity and attributes (pose and expression) of faces through the adversarial training between the generator and discriminator, such that we can conveniently recombine different identities and attributes to synthesize a new facial image x'. Additionally, we can exhibit the smooth transition between different facial expressions through geometry interpolation. (2) Train a pose-invariant FER classifier with the generated image x' and the input x. (3) Retain the identity representation with an identity preservation loss.

1) Geneartor G and Discriminator D_{att}: The generative network is a combination of encoder and decoder. With the input facial image, it first exploits five strided convolutional layers to encode it to a latent space $f(x)$, capturing the facial properties that tend to be stable, i.e., the identity features, followed by two fully connected layer and five fractionally-strided convolutional layers. Then, the new facial image conditioned on the target facial landmark is achieved.

The discriminator D_{att} is to distinguish between 'fake' image x' produced by the generator G, and 'real' image x^t that corresponding to the target facial landmarks g^t. It exploits four strided convolutional layers, followed by two fully connected layer, and outputs a scalar representing the probability that the generated facial image x' comes from the real data distribution p_i. The distribution of the generated faces p_t is supposed to be equivalent to the distribution p_i when optimality is reached. Formulaically, the discriminator on attributes disentangling, D_{att} and G with condition g^t (geometry embedding) can be trained by:

$$L_{D_{att}\_G} = E_{x^t\sim p_i, g^t\sim p_y}[\log D_{att}(x^t, g^t)] + E_{x'\sim p_t, g^t\sim p_y}[\log(1 - D_{att}(x', g^t)], \quad (3)$$

where p_i, p_y, and p_t indicate the distribution of real facial images, real facial landmarks, and generated facial images, respectively. x' represents $G_{dec}(G_{enc}(x), E_{enc}(g^t))$.

2) Geneartor G and Discriminator D_i: The discriminator D_i imposes the uniform distribution f^* on the identity representation $f(x)$, which can help to smooth the face geometry transformation, such as the expression and pose. The discriminator D_i is comprised of four fully connected layer with batch normalization and ReLU non-linearity activation, and the image loss is then formulated as:

$$L_{Di\_G} = E_{f^*\sim\ prior(f)}[\log D_i(f^*)] + E_{x\sim p_i}[\log(1 - D_i(G_{enc}(x)))], \quad (4)$$

where $prior(f)$ is a prior uniform distribution, and $f^* \sim prior(f)$ denotes the random sampling process from $prior(f)$.

3) Classifier C_{exp}: The classifier C_{exp} is a task-specific loss. From the perspective of facial image generation, it is required that the generated image x' should have the new facial expression and pose corresponding to the target facial landmark g^t. Thus, it can be used to penalize the generator loss, which is explicitly helpful for improving the performance of the original generator G. From the perspective of classification, it attempts to classify the expression. Besides, in order to reduce the negative effect caused by the noise in the generated facial images, we introduce a label smoothing regularization [67] trick to the classification loss function. The label smoothing regularization could take the distribution of the non-groundtruth classes into account and encourage the network not to be too confident towards the groundtruth. We adopt a Resnet50 structure [68] for our classification model, and the softmax cross-entropy loss can be defined as:

$$\begin{aligned} L_c(G, C) = E_{x,y^e}[-y^e \log C(x, y^e)] \\ -F_s(y^t) \log C(x', F_s(y^t)), \end{aligned} \tag{5}$$

with

$$F_s(k) = \begin{cases} 1 - \dfrac{(K-1)\rho}{K}, & k = y^t \\ \dfrac{\rho}{K}, & k \neq y^t, \end{cases} \tag{6}$$

where $\rho \sim [0, 1]$ is a hyperparameter, and K means the number of the classes. If ρ is zero, Eq. 5 is just the typical softmax cross-entropy loss. If ρ is too large, the model may fail to predict the groundtruth label. In our case, we follow the definition in [67], and set ρ to 0.1.

4) Identity Preserving Loss: Preserving the identity while synthesizing a new facial image with different expression and pose is a critical part in developing our facial expression recognition model. Therefore, we incorporate the associated constraint by measuring the input-output distance, and the identity loss is then formulated as:

$$L_{con}(G) = L(x - x'), \tag{7}$$

where $L(:, :)$ denotes the ℓ_1 norm.

5) Total Variation Regularization: Usually, the images synthesized by GAN model have many unfavorable artifacts [69], which deteriorate the visualization and the recognition performance. We impose a total regularization term [70] on the final synthesized images to alleviate this issue:

$$L_{TV} = \sum_{c=1}^{C} \sum_{w,h=1}^{W,H} |x'_{w+1,h,c} - x'_{w,h,c}| + |x'_{w,h+1,c} - x'_{w,h,c}|, \tag{8}$$

where W and H represent the width and height of the final synthesized facial image, and C is the number of image channel.

6) Geometry Reconstruction Loss: Reconstruction learning enables latent vector $f(g^t)$ to preserve enough information for the reconstruction of inputs. Thus, for facial landmarks, we employ ℓ_2 distance to compute the difference between reconstructed landmarks and the input landmarks, which is formulated as below:

$$L_{gr} = ||g^t - \widehat{g}^t||, \tag{9}$$

where $\widehat{g}^t$ means the reconstructed landmarks.

7) Overall Objective Function: The final objective loss function is a weighted sum of all the losses defined above:

$$\begin{aligned} \min_{G,C,E} \max_{D} L = \lambda_1 L_{D_{att}\_G} + \lambda_2 L_{D_i\_G} + \lambda_3 L_c \\ + L_{con} + \lambda_4 L_{gr} + \lambda_5 L_{TV}, \end{aligned} \tag{10}$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the trade-off parameters. Sequentially updating this min-max problem makes our generator synthesize face images with various expression and pose guided by geometry embeddings, and further facilitate the learning of our pose-invariant FER model.

C. Discussion

In this section, we show the differences of the proposed model with six most relevant GAN models as shown in Figure 3. The GAN [40] (Figure 3 (a)) and Conditional GAN [45] (Figure 3 (b)) have been introduced in related work. Here, we compare our model with other four most relevant GAN models including Adversarial Autoencoder (AAE) [71] (Figure 3 (c)), a geometry-contrastive-GAN (GC-GAN) [72] (Figure 3 (d)), conditional adversarial autoencoder (CAAE) [66] (Figure 3 (e)), and disentangled representation learning-GAN (DR-GAN) [73] (Figure 3 (f)). The details are as follows.

1). As shown in Figure 3 (c), the AAE has two objectives in order to turn an autoencoder into a generative model: the autoencoder reconstructs the input image, and the latent vector generated by the encoder matches an arbitrary prior distribution by training D. Different from the AAE, our model (Figure 3 (g)) can explicitly disentangle the identity representation learning from both expression and pose variations by use of the corresponding geometry embedding.

2). The GC-GAN [72] is an extension of the conditional GANs, which is shown in Figure 3 (d). Through injecting geometry information into the conditional GANs, the GC-GAN could synthesize face images with new expressions, which is the method most related to our proposed model. However, the GC-GAN is used only for facial image generation. Different from this method, we incorporate a classifier into our model, which is helpful for expression classification and can improve generator performance through penalizing the generator loss. Furthermore, we can smooth the geometry transformation through imposing uniform distribution on the identity representation by using an extra discriminator.

3). As shown in Figure 3 (e), the CAAE [66] extends the AAE to generate face images with different ages. In this method, the age is incorporated with a one-hot vector. Different from this method, our model uses the facial landmarks as geometry guided information in facial image synthesis. Thus, our model is more flexible and

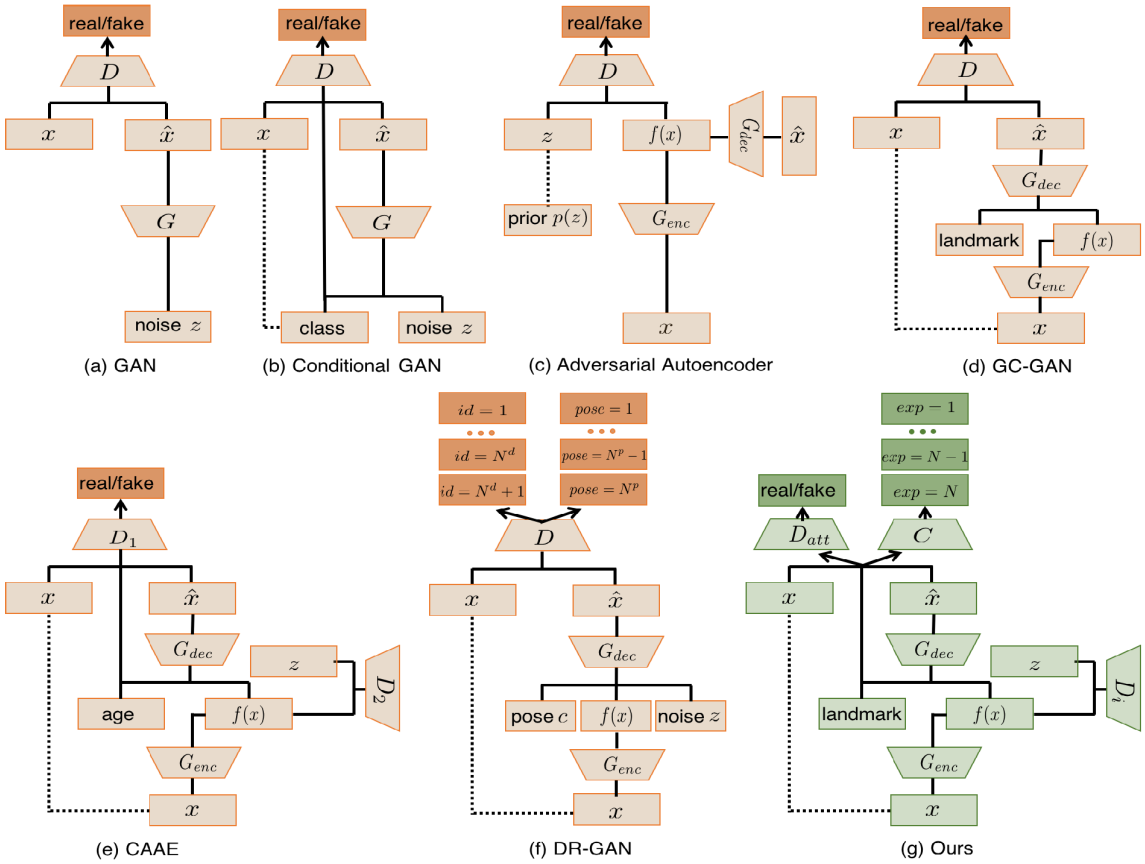


Fig. 3. Illustration of the structure of GAN [40], Conditional GAN [45], Adversarial Autoencoder (AAE) [71], GC-GAN [72], CAAE [66], DR-GAN [73], and the proposed method. Where x and $\hat{x}$ are the input and generated image. $G(/G_*)$, $D(/D_*)$, C are generator, discriminator, and classification network, respectively. z is the noise vector sampled from a prior distribution.

general, which can be applied to the smooth transition between different facial expressions through geometry interpolation. Moreover, our model embeds a classifier in the network and can strive for the generated facial image to have the target expression that we need.

4). The DR-GAN [73], as shown in Figure 3 (f), generalizes the GAN to learn a discriminative classifier which is trained to not only distinguish between real and fake images, but also classify real images into N^d and N^p classes. It is a variational autoencoder-based method to learn disentangled representation for face recognition task. Different from the DR-GAN, the proposed model is mainly for generating more labeled facial images to train a deep network classifier for the FER, because the training samples are the main bottleneck in facial expression recognition. Furthermore, we disentangle both the expression and pose from the facial images by the facial landmark, and introduce a separated classifier for the FER task.

IV. Experimental Results

In this section, we will first clarify the datasets used in our method (Sec. IV-A) and the implementation of the proposed method (Sec. IV-B). Then we will show experimental results of our model for facial image synthesis (Sec. IV-E) and pose-invariant FER (Sec. IV-C). For the former task, we show qualitative results of the generated facial images under different poses and expressions. We also show examples about the smooth transitions of different facial expressions. For the latter one, we quantitatively evaluate the expression recognition performance using the generated and original facial images. We first conduct extensive experiments to compare our method with the state-of-the-arts, and then conduct ablation study to demonstrate the effectiveness of all components in our model.

A. Datasets

To demonstrate the effectiveness of the proposed model, we conduct extensive experiments on three standard datasets including (1) Multi-PIE [26]: the public multi-pose facial expression dataset, (2) BU-3DFE [27]: the 3D facial expression dataset, and (3) SFEW [25]: the static facial expressions in the wild dataset. The details are as follows.

1) Multi-PIE: The Multi-PIE is for evaluating facial expression recognition under pose and illumination variations in the controlled setting. Following the setting in [10], we use images of 270 subjects depicting acted facial expressions of Neutral (NE), Disgust (DI), Surprise (SU), Smile (SM), Scream(SC), and Squint (SQ), captured at five pan angles $-30°$, $-15°$, $0°$, $15°$ and $30°$, resulted in 1531 images per pose. Consequently, we have $1,531 \times 5 = 7,655$ facial images in total for

our experiments. We perform five-fold subject independent cross-validation on the Multi-PIE. As a result, the training dataset comprises 6, 124 facial images whereas the testing one comprises 1, 531 facial images.

2) BU-3DFE: The BU-3DFE is a 3D facial expression dataset having 100 subjects with 3D models and facial images. It contains images depicting seven facial expressions: Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU) and Neutral (NE). With the exception of the neutral expression, each of the six prototypic expressions includes four levels of intensity. Following the setting in [17], [52], [74], [75], we render 2D facial images from the 3D models at the fourth level of intensity, six universal facial expressions (AN, DI, FE, HA, SA, SU), and 35 poses including 7 pan angles ($0°, \pm 15°, \pm 30°, \pm 45°$), and 5 tilt angles ($0°, \pm 15°, \pm 30°$)). Consequently, we have $100 \times 6 \times 35 \times 1 = 21,000$ face images in total for our experiments. We randomly divide the 100 subjects into a training set with 80 subjects and a testing one with 20 subjects, such that there are no overlaps between the training subjects and the testing subjects. As a result, the training set comprises 16, 800 facial images whereas the testing one comprises 4, 200 facial images.

3) SFEW: The SFEW is a dataset in the wild with 95 subjects. It consists of 700 images (346 images in Set 1, 354 images in Set 2) extracted from movies covering unconstrained facial expressions, varied head poses, changed illumination, large age range, different face resolutions, occlusions, and varied focus. The images are labeled with Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU) and Neutral (NE). In our experiment, we consider all the seven facial expressions, and adopt cross dataset experiments on it. Specifically, we train the facial expression classification model on the BU3DFE, and test it on the SFEW.

B. Network Architectures and Settings

We construct the network according to Figure 2. Prior to feeding the images into the networks, we first use the lib face detection algorithm to get 68 facial landmarks and crop out the faces [65], and then resize them as 224×224. After that, the image intensities and the facial landmarks are linearly scaled to the range of $[-1, 1]$, which are the original inputs of our model. Normalizing the input may make the training process converge faster. As shown in Figure 2, our model is comprised of five basic subnetworks, i.e., a facial geometry embedding network E, a generator G, two discriminators D_i and D_{att}, and a classifier C_{exp}. In addition to the classifier C_{exp}, which we just employ a typical ResNet50 [68] by adding two more fully-connected layers after $pool5$, the detailed structures of the other modules are provided in Table II. The geometry embedding network E employs an encoder-decoder structure with six fully connected layers. To stabilize the training process, the designing of the network architectures of G, D_{att}, and D_i are based on the techniques in the CAAE [66]. Specifically, as shown in Table II, the generator network G is a convolutional neural network without batch normalization. The G_{enc} has five convolution layers without batch normalization,

TABLE II

THE DETAILED ARCHITECTURE OF OUR PROPOSED MODEL. IN $((De)Conv(d, k, s))$, d, k, AND s STAND FOR THE NUMBER OF FILTERS, THE KERNEL SIZE, AND THE STRIDE, RESPECTIVELY. FC REPRESENTS A FULLY-CONNECTED LAYER. BN IS BATCH NORMALIZATION. $LReLU$ REFERS TO $leakyReLU$

Geometry Embedding E	**Generator** G
FC(128), BN, ReLU	Conv(64, 5, 2), ReLU
FC(64), BN, ReLU	Conv(128, 5, 2), ReLU
FC(32), BN, ReLU	Conv(256, 5, 2), ReLU
FC(64), BN, ReLU	Conv(512, 5, 2), ReLU
FC(128), BN, ReLU	Conv(1024, 5, 2), ReLU
FC(136), Tanh	FC(50), Tanh
Discriminator D_i	Concat(82)
FC(64), BN, ReLU	FC(50176), ReLU
FC(32), BN, ReLU	DeConv(1024, 5, 2), ReLU
FC(16), BN, ReLU	DeConv(512, 5, 2), ReLU
FC(1)	DeConv(256, 5, 2), ReLU
Discriminator D_{att}	DeConv(128, 5, 2), ReLU
Conv(16, 5, 2), BN, ReLU	DeConv(64, 5, 2), ReLU
Conv(32, 5, 2), BN, ReLU	DeConv(32, 5, 2), ReLU
Conv(64, 5, 2), BN, ReLU	DeConv(16, 5, 2), Tanh
Conv(128, 5, 2), BN, ReLU	
FC(1024), BN, LReLU	
FC(1)	

and a fully connected layer with tangent activation function. The output is the identity representation $f(x)$, which is then concatenated with the desired facial landmark embedding to construct the input of G_{dec}. The G_{dec} is constructed by one fully connected layer and seven fractionally-strided convolution layers to transform the concatenated vector into a synthetic image x', which is the same size as the input x. The feature intensities of the synthetic facial image are also the range of $[-1, 1]$ through hyperbolic tangent function. In the discriminators D_{att} and D_i, the batch normalization is applied after each convolution layer to stable the model training.

In the training phase, facial landmarks are extracted from the corresponding target images (the same identity with the input), which is helpful for disentangling the identity representation from the expression and pose variations. After the training, the proposed model can automatically generate labeled facial images with arbitrary expressions and poses. Then, during the testing (of the GAN), we just extract facial landmarks from 60 arbitrary facial images under different expressions and poses, which can increase the intra-class distance in the generated new facial images. The coefficients λ_1, λ_2, λ_3, λ_4 and λ_5 in the loss L are empirically set to 10^{-3}, 10^{-3}, 10^{-3}, 10^{-4} and 10^{-4}, respectively. The model is implemented by using TensorFlow [76] and is trained with the ADAM optimizer [77], which is used with a learning rate of 5^{-4}, beta1 of 0.5, and beta2 of 0.999. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.02. More details about the network architectures and settings could be found in our source code, which is included in the supplementary material.

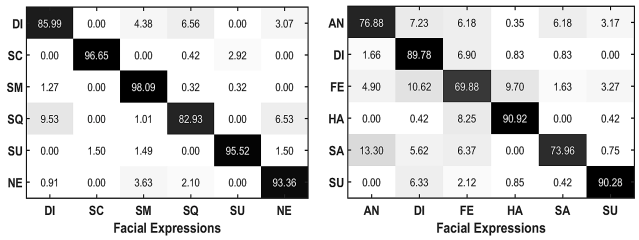
C. Quantitative Results

1) Experiments on the Multi-PIE Dataset: In this section, we report the recognition accuracy on Multi-PIE in two cases: First, detailed recognition accuracy on Multi-PIE.

TABLE III

RESULTS ON THE MULTI-PIE DATASET IN TERMS OF THE RECOGNITION RATES (%). THE LEFTMOST COLUMN INDICATES DIFFERENT VIEWS, AND THE TOP ROW INDICATES DIFFERENT FACIAL EXPRESSIONS. THE HIGHEST ACCURACY IS HIGHLIGHTED IN BOLD

Pose / Exp.	DI	SC	SM	SQ	SU	NE	Ave.
-30	84.78	95.74	100	82.05	97.44	95.45	92.58
-15	89.13	100	98.41	87.50	95.00	92.42	93.74
$+0$	77.78	97.92	98.44	75.61	92.68	95.45	89.65
$+15$	89.13	89.58	96.72	79.49	92.50	92.42	89.97
$+30$	89.13	100	96.88	90.00	100	91.04	**94.51**
Average	85.99	96.65	**98.09**	82.93	95.52	93.36	**92.09**



(a) On the Multi-PIE dataset. (b) On the BU-3DFE dataset.

Fig. 4. The average confusion matrix. The average recognition rate is 92.09% and 81.95%, respectively.

Second, comparison with several state-of-the-art facial expression recognition methods.

Table III shows the recognition accuracy on this dataset. The rightmost column represents the average recognition error rates for different views (a total of 5 views). The bottom row represents the average recognition error rates for different facial expressions (a total of 6 universal facial expressions), and the bottom-right corner cell represents the average overall recognition error rate. As it is shown, the three expressions of smile, scream and surprise are much easier to be recognized than others, which is most likely because of their relatively large muscle deformations. For the smile and surprise expressions, all the accuracies of the five yaw head poses are higher than 92%. Followed are the recognition results of disgust and neutral, which are more than 85%. The lowest accuracy is 82.93% of the expression squint. The confusion matrix on this dataset is shown in Figure 4(a), from which we can see that the main error classification comes from the confusion between disgust and squint. There are 6.56% of disgust samples misclassified to be squint and 9.53% of quint samples misclassified to be disgust. The high confusion may be caused by the fact that the expressions of disgust and squint have similar muscle deformations around eyes, which is consistent with some previous works in facial expression recognition, such as [10].

We then compare the FER recognition accuracy of the proposed model with the results achieved by various algorithms on Multi-PIE, including the algorithms based on hand-crafted features and deep learning methods. **For the algorithms based on hand-crafted features**, we compare the results reported in [10] including kNN, LDA, LPP, D-GPLVM, GPLRF, GMLDA, GMLPP, MvDA, and DS-GPLVM. The detailed results across all views are summarized in Table IV. The mean FER accuracy is reported in the last column. It can

TABLE IV

COMPARISON OF STATE-OF-THE-ART METHODS ON THE MULTI-PIE DATASET. THE HIGHEST ACCURACY FOR EACH POSE IS HIGHLIGHTED IN BOLD

Methods	Poses					Average
	-30	-15	0	15	30	
kNN	80.88	81.74	68.36	75.03	74.78	76.15
LDA	92.52	94.37	77.21	87.07	87.47	87.72
LPP	92.42	94.56	77.33	87.06	87.68	87.81
D-GPLVM	91.65	93.51	78.70	85.96	86.04	87.17
GPLRF	91.65	93.77	77.59	85.66	86.01	86.93
GMLDA	90.47	94.18	76.60	86.64	85.72	86.72
GMLPP	91.86	94.13	78.16	87.22	87.36	87.74
MvDA	92.49	94.22	77.51	87.10	87.89	87.84
DS-GPLVM	**93.55**	**96.96**	82.42	89.97	90.11	90.60
VGG16	82.95	84.33	80.67	83.11	86.23	83.46
VGG19	84.28	85.15	82.67	84.54	86.93	84.71
ResNet50	87.54	87.71	84.21	85.90	87.54	86.58
DenseNet121	87.71	87.88	84.54	86.23	86.89	86.65
Ours	92.58	93.74	**89.65**	**89.97**	**94.51**	**92.09**

be seen that our method owns the best recognition accuracy on the testset reaching 92.09%. This result exceeds the others with a satisfactory growth from 1.49% to 15.94% in terms of the FER accuracy. Specifically, although all the models cannot achieve good performances in the frontal view, our method get a significant improvement for it, whose average accuracy increases by 7.23% to 21.29%. Furthermore, the FER accuracy under 30° is also improved, with an increase from 4.40% to 19.73%. For the performance on the other poses, our model could also achieve highly competitive results. The improvement may be caused by the deep learning based method used in our model. The algorithms based on hand-crafted features cannot sufficiently and efficiently cope with the occlusion caused by pose variation [78]. **For the deep learning methods**, we compare the results with the networks including VGG16 [79], VGG19 [79], ResNet50 [68], and DesNet121 [80]. The weights of these methods are initialized by 'imagenet', and then fine-tuned by the same training set as our method. In order to apply these methods to our task, we omit the final fully connected layer of them, and use the features from the last polling layer to train the facial expression classification model. The lower part of Table IV shows the performance of different deep learning based methods. Our method also achieves the highest FER results. Such improvement indicates that our facial expression recognition task can get benefits from the generated images with arbitrary poses and expressions.

2) Experiments on the BU-3DFE Dataset: We further evaluate the FER performance of the proposed method on BU-3DFE. Following the setting on Multi-PIE dataset, we report the recognition accuracy on BU-3DFE in two cases as well: First, detailed recognition accuracy on BU-3DFE. Second, comparison with several state-of-the-art facial expression recognition methods.

The detailed experimental results over each expression and pose are shown in Table V. 35 poses and 6 facial expressions are used in our experiment. The rightmost column represents the average recognition error rates for 35 different poses, and the bottom row represents the average recognition error rates for 6 different facial expressions. The bottom-right

TABLE V

RESULTS ON THE BU-3DFE DATASET IN TERMS OF THE RECOGNITION RATES (%). THE LEFTMOST COLUMN INDICATES DIFFERENT VIEWS (PAN AND TILT ANGLES x, y IN DEGREES), AND THE TOP ROW INDICATES DIFFERENT FACIAL EXPRESSIONS. THE HIGHEST ACCURACY IS HIGHLIGHTED IN BOLD

Pose / Exp.	AN	DI	FE	HA	SA	SU	Ave.
$-45, -30$	62.50	75.00	77.78	87.50	66.67	75.00	74.07
$-45, -15$	58.33	87.50	66.67	100	75.00	88.89	79.40
$-45, +0$	80.00	100	62.50	100	66.67	85.71	82.48
$-45, +15$	100	87.50	60.00	100	75.00	88.89	85.23
$-45, +30$	72.73	75.00	71.43	83.33	100	100	83.75
$-30, -30$	71.43	100	58.33	88.89	83.33	100	83.66
$-30, -15$	60.00	90.00	81.82	88.89	83.33	100	84.01
$-30, +0$	83.33	66.67	57.14	85.71	77.78	85.71	76.06
$-30, +15$	88.89	85.71	57.14	88.89	66.67	100	81.22
$-30, +30$	88.89	87.50	60.00	85.71	71.43	88.89	80.40
$-15, -30$	62.50	75.00	83.33	100	50.00	85.71	76.09
$-15, -15$	75.00	100	60.00	100	66.67	87.50	81.53
$-15, +0$	75.00	83.33	83.33	75.00	66.67	100	80.56
$-15, +15$	85.71	87.50	63.64	83.33	54.55	75.00	74.95
$-15, +30$	71.43	100.00	66.67	83.33	87.50	83.33	82.04
$+0, -30$	75.00	75.00	83.33	100	62.50	80.00	79.31
$+0, -15$	66.67	87.50	62.50	100	80.00	100	82.78
$+0, +0$	77.78	85.71	80.00	87.50	73.33	100	84.05
$+0, +15$	77.78	85.71	100	100	80.00	100	**90.58**
$+0, +30$	100	100	63.64	83.33	66.67	100	85.61
$+15, -30$	83.33	100	85.71	88.89	100	83.33	90.21
$+15, -15$	62.50	100	50.00	90.00	80.00	87.50	78.33
$+15, +0$	100	87.50	100	100	71.43	80.00	89.82
$+15, +15$	100	100	42.86	75.00	100	100	86.31
$+15, +30$	70.00	80.00	71.43	100	72.73	87.50	80.28
$+30, -30$	66.67	100	85.71	100	83.33	100	89.29
$+30, -15$	85.71	83.33	57.14	100	63.64	85.71	79.26
$+30, +0$	75.00	100	85.71	100	75.00	90.91	87.77
$+30, +15$	84.62	87.50	75.00	83.33	75.00	85.71	81.86
$+30, +30$	75.00	91.67	60.00	100	66.67	85.71	79.84
$+45, -30$	71.43	100	80.00	84.62	57.14	85.71	79.82
$+45, -15$	66.67	100	71.43	87.50	75.00	85.71	81.05
$+45, +0$	63.64	100	71.43	75.00	85.71	100	82.63
$+45, +15$	71.43	100	60.00	87.50	66.67	85.71	78.55
$+45, +30$	81.82	77.78	50.00	88.89	62.50	91.67	75.44
Average	76.88	89.78	69.88	**90.92**	73.96	90.28	**81.95**

corner cell represents the average overall recognition error rate. The results show that our method obtains the average recognition accuracy of 81.95%. Furthermore, among the six expressions, happiness and surprise are easier to be recognized with accuracy over 90%. This is most likely due to the fact that the muscle deformations of both expressions are relatively large compared with others. Moreover, fear is the most difficult expression to be recognized, with the lowest at 69.88%, followed by sadness. These are also reflected in the confusion matrix in Figure 4(b). Note that a main contributing factor to the poor performance of fear is its confusion with happiness and disgust. This coincides with the findings of Moore and Bowden in [14], where the authors point out that the confusion is due to the expressions of fear and happiness having similar muscle deformation around the mouth. The confusion between fear and disgust is probably because they have similar muscle deformation around the mouth and nose. In addition, another two expressions likely to be confused are sadness and anger. These two expressions have the least amount of facial movement and thus are difficult to distinguish.

Then, the performance of our method is compared to seven well-established algorithms based on hand-crafted features, which are reported in [11], [14], [17], [52], [57], [74], [81], and five deep learning based methods, namely, VGG16 [79], VGG19 [79], ResNet50 [68], DesNet121 [80], and a method

TABLE VI

COMPARISON OF THE AVERAGE RECOGNITION ACCURACY WITH STATE-OF-THE-ART METHODS FOR THE FER ON THE BU-3DFE DATASET

Methods	Poses			Ave.
	tilt	pan	total	
Zheng 2014 [57]	-	$(0^\circ, +90^\circ)$	5	66.0
Moore and Bowden 2011[14]	-	$(0^\circ, +90^\circ)$	5	71.1
Zhang et al. 2009 [81]	-	$(0^\circ, +90^\circ)$	5	78.3
Zheng 2014 [57]	-	$(0^\circ, +90^\circ)$	5	78.9
Zhang et al. 2016 [11]	-	$(0^\circ, +90^\circ)$	5	80.1
Tang et al. 2010 [17]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	75.3
Tariq et al. 2013 [74]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	76.34
Tariq et al. 2014 [53]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	76.60
VGG16	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	73.03
VGG19	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	75.11
ResNet50	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	77.41
DenseNet121	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	77.69
Jampour et al. 2015 [75]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	78.64
Ours	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	81.95

reported in the literature [75]. Specifically, methods [14], [57], [75], methods [17], [52], [57], [74] and [81] train their models based on the hand-crafted features LBP, SIFT, and geometry features (83 landmark points), respectively. The VGG16 [79], VGG19 [79], ResNet50 [68], and DesNet121 [80], as mentioned on the Multi-PIE dataset, are initialized by 'imagenet', and then fine-tuned by the same training set as our method. In [75], the SIFT feature is used as the input of DNN to learn features. Here, the model is trained separately for each step. Different from this method, ours is an end-to-end learning model. Details regarding each reported results are summarized in Table VI. The proposed method is evaluated under the same experiment setting with them, and the average FER accuracy over the expressions is reported in the last column. From Table VI, we can conclude the following: **For the algorithms based on hand-crafted features**, although the methods in [11], [14], [57], [81] conduct the FER on a relatively small set of discrete poses just containing 5 pan angles, our method is also competitive to the results achieved by these methods with a 1.85% to 15.95% improvement on the FER accuracy. In addition, the algorithms in [17], [52], [74] use the facial images with 35 poses to train their model, which are the same as ours. Generally, our method still consistently outperforms the state-of-the-art methods, which is higher than the runner-up (Tariq et al. [52]) with a 5.35% gap. The gains over these methods may contribute to the deep learning based methods adopted in our method, which can better deal with the nonlinear facial texture warping caused by pose and individual difference. **For the deep learning based methods**, from the lower part of table VI, we can see that the proposed method also achieves the best accuracy (3.31% to 8.92% higher than others). It may because that far less labeled data is typically available for training such emotion classification systems, thus automatically generating sufficient labeled images is essential to the task.

3) Cross Dataset Experiments on the BU-3DFE and SFEW Datasets: Finally, we evaluate the ability of our proposed method to generalize to unseen real-world data on the SFEW dataset. Specifically, first, the images on the BU-3DFE dataset are used for training. Then, images from the SFEW are

TABLE VII

COMPARISON OF THE AVERAGE RECOGNITION ACCURACY (%) WITH STATE-OF-THE-ART METHODS ON THE SFEW DATASET. THE HIGHEST ACCURACY FOR EACH EXPRESSION IS HIGHLIGHTED IN BOLD

Method / Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Average
Baseline	23.00	13.00	13.90	29.00	23.00	17.00	13.50	18.90
MvDA	23.21	17.65	27.27	40.35	**27.00**	10.10	13.19	22.70
GMLDA	23.21	17.65	**29.29**	21.93	25.00	11.11	10.99	19.99
GMLPP	16.07	21.18	27.27	39.47	20.00	19.19	16.48	22.80
DS-GPLVM	25.89	**28.24**	17.17	42.98	14.00	**33.33**	10.99	24.70
Ours	**29.09**	24.88	17.65	**51.19**	20.00	29.20	**18.70**	**27.24**

TABLE VIII

EFFECT OF DIFFERENT COMPONENTS ON FACIAL EXPRESSION CLASSIFICATION ACCURACY (%)

Methods	Poses					Average
	-30	-15	0	15	30	
No_E2E	90.37	90.97	87.54	87.00	92.13	89.55
No_LS	90.70	91.36	88.00	88.20	92.46	90.14
No_SynFI	87.54	87.71	84.21	85.90	87.54	86.58
Ours	92.58	93.74	89.65	89.97	94.51	92.09

used for testing. Apparently, it is a rather challenging task mainly because the test images are captured in an uncontrolled environment, which is characterized by large variation in head-poses, illumination, and occlusions of parts of the face. However, our model is trained using data of deliberated expressions under the laboratory conditions, which can differ considerably in subtlety compared to the spontaneous expressions used for testing.

Table VII provides the comparison results of our method with the current state-of-the-art results [10], [82], [83] including MvDA, GMLDA,GMLPP, DS-PLVM, and the baseline designed by the dataset creators [25] on the SFEW dataset. The average recognition accuracies across all the expressions of each method are reported in Table VII. From the average FER accuracy reported in the last column of this table, we can see that all of the methods cannot achieve good performance on this dataset, especially in the surprise expression. However, the proposed model can significantly improve the recognition accuracy in this expression, and our method can achieve the highest average recognitionn accuracy of 27.24%, which outperforms all existing methods with a 2.54% to 8.34% improvement. This may attribute to the generated facial images, which can train a well classification model with sufficient samples.

D. Component Analysis of the Proposed Model

1) Model Analysis: To help analyze the proposed model and show the benefit of each module, we design several ablated versions of our method. Specifically, we train three variants of our method:

a). **No synthesized facial images (No_SynFI)** omits the synthesized facial images when training the classification model, which means that the FER classifier is only trained by the original facial images on the dataset.

b). **No label smooth (No_LS)** omits the label smooth trick used in Eq. 5, which means that the generated facial images are used equivalent to the original facial images when training the facial expression classification model.

TABLE IX

FER ACCURACY WITH DIFFERENT LEVELS OF NOISE ($[-\sigma, \sigma]$, $\sigma = 2, 4, 6, 8$ PIXELS) ADDED TO LANDMARK LOCATIONS

σ	-30	-15	0	15	30	Average
$\sigma = 2$	90.64	93.09	89.11	90.76	94.72	91.88
$\sigma = 4$	89.63	91.09	87.13	90.76	92.43	90.21
$\sigma = 6$	85.28	86.84	83.17	84.49	88.12	85.58
$\sigma = 8$	82.61	82.51	79.87	82.51	83.88	82.27
$\sigma = 0$ (Ours)	92.58	93.74	89.65	89.97	94.51	92.09

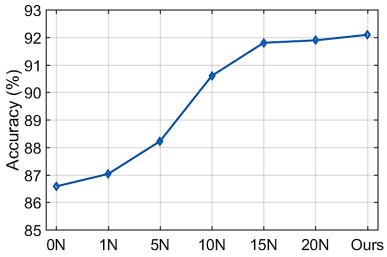


Fig. 5. Effect of the number of training samples for the FER task.

TABLE X

COMPARISON TO ONE-SHOT LEARNING METHOD ON MULTI-PIE DATASET

Methods	DI	SC	SM	SQ	SU	NE	Ave.
$\text{MSN}_{DI\&SC}$	44.44	60.92	91.17	70.35	91.00	95.69	74.71
$\text{MSN}_{SM\&SQ}$	77.78	92.86	48.90	37.69	90.00	89.19	72.73
$\text{MSN}_{SU\&NE}$	78.22	90.34	92.11	68.84	54.50	47.15	71.86
Average	85.99	96.65	98.09	82.93	95.52	93.36	92.09

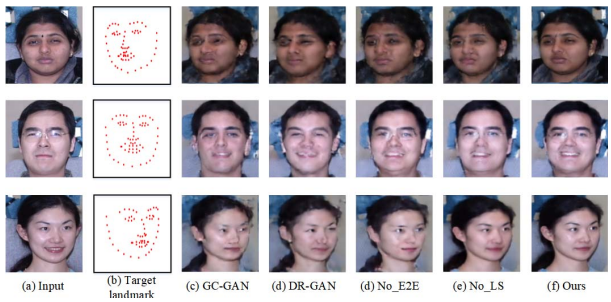


Fig. 6. Synthesis results of different methods. Each row from left to right are the real samples, target landmarks, and results generated by GC-GAN, DR-GAN, No_E2E, No_LS and our model, respectively.

c). **No end-to-end (No_E2E)** omits the end-to-end strategy used in our method, so the facial image synthesis part and facial expression recognition part are trained separately.

Table VIII shows the performance of each setting. We observe that the proposed method is obviously better than its variants across all poses. Without the **No_E2E**, the performance drops by 2.54% on the average FER accuracy, highlighting the importance of the designed end-to-end strategy, which could facilitate the free parameters in the facial image generation and classification becoming co-adapt and cooperate, and thus promoting the FER task. Without the **No_LS**, the FER accuracy also shows a light drop, about 1.95%. Particularly, when we omit the **No_SynFI**, the accuracy drops significantly for all poses, this observation naturally

(a) Input with -30° and SM.

(b) Input with -30° and NE.

(c) Input with 0° and SC.

(d) Input with 0° and SQ.

Fig. 7. Example results of the generated facial images with different poses and expressions via the proposed model. Each row contains samples generated with different poses. And each column contains samples generated with different expressions. The orange rectangle: the input image. The yellow rectangle: the ground truth image corresponding to the input. The green rectangle: the generated facial images.

evidences that sufficient training samples are essential to the deep learning based method.

2) Effect of Landmark Localization Noise on FER: In this experiment, landmark locations of the facial images used in image synthesis network (during testing) are deliberately corrupted by different levels of noise (randomly selected from the interval) in $[-\sigma, \sigma]$, with $\sigma = 2, 4, 6, 8$ pixels. The mean FER accuracies with different intervals are reported in Table IX. In each row of Table IX, the results are achieved on the data corrupted by the same interval of noise, and averaged over all expressions for each pose. The last column of Table IX shows the accuracy averaged FER accuracy for each interval of the noise. Clearly, the clean data ($\sigma = 0$) can always get the highest FER accuracy. When the data is corrupted by the noise, there is an expected decline in the performance over each poses. However, the proposed method remains stable with low levels of noise ($\sigma = 2, 4$). When σ is larger than 6, the mean FER accuracy consistently degenerates. We observe the generation results, and find that the generated facial images lost their true expression information with the high level noise. This, in turn, results in negative influence on the FER accuracies. Generally, our model is quite tolerable regarding landmark localization errors.

3) Effect of the Generated Facial Images on FER: In order to further verify the effect of the generated facial images, we compare our method with the models trained with different number of generated images. Given the original N images, we can obtain $5 \times 6 \times N$ generated images. To evaluate the effect of the training data size, we randomly choose $0 \times N$, $1 \times N$, $5 \times N$, $10 \times N$, $15 \times N$, $20 \times N$ images from the generated facial images during each training epoch, and then incorporate them with the original images to train the classifier, where $0 \times N$ means that the classifier is trained only using the original images. Specifically, we denote them as $0N$, $1N$, $5N$, $10N$, $15N$, $20N$. Figure 5 shows the average classification rate of the proposed method with different number of training

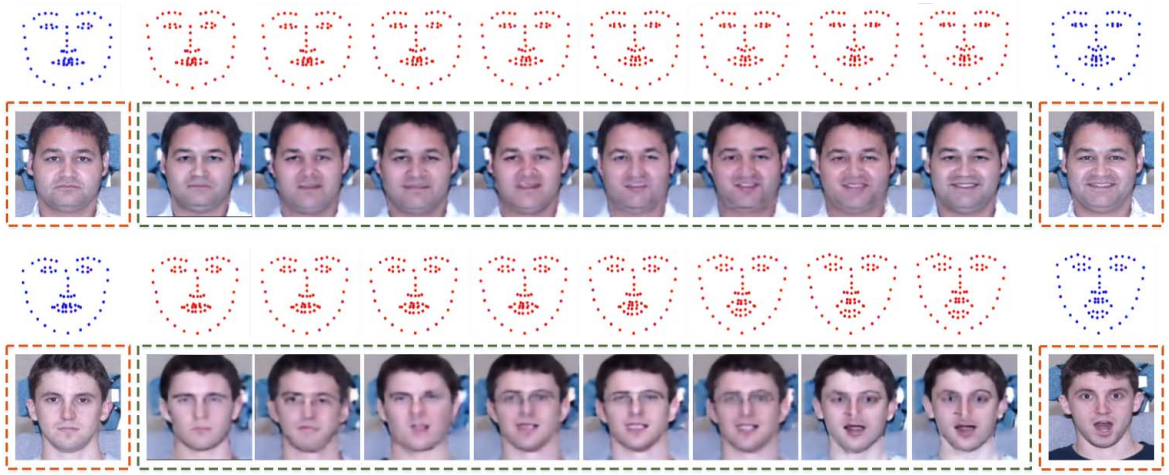


Fig. 8. Smooth transitions of facial expressions achieved by geometry interpolation from one facial expression to the other. The first and third rows show the progressive changes of the facial landmarks. And the second and fourth rows show the facial expression transitions from neutral to happiness, and from neutral to surprise, respectively.

samples. It is clear that our model achieves the best recognition results. Besides, we can also find out that the average accuracy of the FER can be improved with the increase of the number of training samples, which further indicates the importance of generating more labeled training samples.

Finally, we compare our proposed model with one-shot learning method, which is also usually used to handle the data insufficient problem. As we did not find existing one-shot FER results, we reimplemented the method in [84] for comparison,[1] which is composed of a convolutional siamese network. Specifically, we use the Multi-PIE dataset for verification, and divide it into two parts - 4 expressions are used for training and validation, while the remaining 2 expressions are test classes for the one-shot task. As the original architecture of the method in [84] only contains 5 convolutional layers and a final global pooling layer, which can not deal well with the complex facial images according to our experimental results, thus we modify the basic architecture in it to Resnet [68]. The modified siamese network is named as $\text{MSN}_{*\&*}$ as shown in Table X, where $*$ means the test expressions. As can be seen from Table X, although the one-shot learning method can get relatively high accuracy on the training expressions, it is hard to generalized to the test ones. Overall, our model achieves the highest accuracy, and with a significant improvement in classifying images of the test expressions. We can conclude as follows. First, the full supervised learning method can usually get promising results. Second, generating more labeled training data is beneficial to the classification task.

E. Qualitative Results

1) Facial Image Synthesis: In order to evaluate whether the synthetic facial images are generated conditioned on target expression, we qualitatively compare the generated faces with the ground truth. The qualitative results are shown in Figure 7. We randomly select a facial image from the test set, which is shown in the orange rectangle in each subfigure. We aim to translate this facial image into the new faces with different poses (5 poses) and expressions (6 expressions), which are shown in the green rectangle. For easy comparison, the ground truth are shown in the yellow rectangle. From these examples, it is clear that the generated facial images are identity-preserving and the attributes (expression and pose) have been jointly modeled in the identity representation as shown in the red rectangles in Figure 7 (a), indicating the effectiveness of guiding landmarks.

From Figure 7, we can also conclude the following observations: 1) As shown in Figure 7 (a) and (b), the real samples (in the red rectangle) of DI, SC, SQ and SU expression are dressed differently with the test image (in the orange rectangle). However, we can preserve the identity features of the test images while changing the expressions of the person. 2) As the examples illustrated in Figure 7 (c) and (d), the ground truth shown in the red rectangle has the different hairstyle with the test image, especially for the images in Figure 7 (d), in which we can see the person in the test image has the curly hair while the hair is tied up in the ground truth. However, we can also synthesize convincing facial images with the target expressions, which indicate the effectiveness of the proposed method. More detailed qualitative results can be found in the supplementary material.

2) Facial Expression Transfer: This section presents the results of facial expression transfer. The goal of the facial expression transfer is to generate faces between two expressions, which can also be used to evaluate whether there exists semantic correspondence between the identity embedding and the geometry embedding of facial landmarks. Here, we use the geometry traversal for qualitative analysis. Specifically, we would like to seek a sequence N images with a smooth transition between two different facial expressions of the same person (e.g., x^u, x^v). In our case, we conduct a linear interpolation $[\frac{k}{N} \cdot g^{t_u} + (1 - \frac{k}{N} \cdot g^{t_v})]_{k=0}^{N}$ on facial landmarks, in which the g^{t_u} and g^{t_v} are the corresponding facial landmarks

[1]the code is provided by: https://github.com/Goldesel23/Siamese-Networks-for-One-Shot-Learning

for x^u and x^v. Figure 8 shows smooth transitions between different facial expressions. Based on the results in Figure 8, it is clear that each generated image sequence is just like a video recording the dynamic expression changes of a person.

3) Image Synthesis Results Comparison: In Figure 6, we compare our generated facial images with 4 different GANs, including our implementations of GC-GAN [72] and DR-GAN [73], and 2 variants (No_E2E, No_LS) of our proposed method. Specifically, for the GC-GAN and DR-GAN, we get the code from the github,[2] and retrained the model for our task. We replace the pose encoding in DR-GAN with the facial geometry embedding to generate facial images under different expressions and poses. According to the description in Section IV-C, No_E2E and No_LS mean that the end-to-end strategy and the label smooth trick adopted in our model are omitted in these methods. As shown in Figure 6, our method could provide a higher visual quality of image synthesis results compared to the other methods. Specifically, compared to GC-GAN and DR-GAN, our method demonstrates an advantage in image smoothness. We conjecture that this is because our method could impose uniform distribution on the identity representation by using an extra discriminator. While No_E2E and No_LS could synthesize relatively smooth image, the generated results of them are blurry. We believe that the image generation task and the FER are positively correlated. In other words, the generated facial images can facilitate the FER, and the FER can improve the performance of the generator through penalizing the generator loss in turn.

V. Conclusion

In this paper, we present an end-to-end learning model for simultaneous facial images synthesis and pose-invariant facial expression recognition conditioned on the geometry information. Through disentangling the attributes (expression and pose) from the facial image, we can generate facial images with arbitrary expressions and poses to help train the deep neural classification model. Experimental results on three standard datasets demonstrate the effectiveness of our model. Furthermore, our method can also be applied to facial expression transfer. In the future, we attempt to achieve a fully automatic system for facial landmark detection and incorporate it into our model, instead of conducting the facial landmark detection in advance. In addition, the proposed model is general and can be applied to other classification tasks, such as face recognition, image classification, and audio event recognition, which we leave as our future work.

References

[1] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017.

[2] W.-S. Chu, F. D. L. Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[3] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.

[2]The code is provided by: GC-GAN: https://github.com/joffery/GC-GAN; DR-GAN: https://github.com/kayamin/DR-GAN.

[4] G. Peng and S. Wang, "Weakly supervised facial action unit recognition through adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2188–2196.

[5] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.

[6] P. Ekman and W. V. Friesen, *Pictures of Facial Affect*. Palo Alto, CA, USA: Consulting Psychologists Press, 1976.

[7] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. Int. Conf. Smart Comput.*, Nov. 2014, pp. 303–308.

[8] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 19–27.

[9] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812.

[10] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[11] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[12] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[13] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel, "Multi-view facial expression recognition using local appearance features," in *Proc. ICPR*, 2012, pp. 3533–3536.

[14] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 4, pp. 541–558, Apr. 2011.

[15] Z. Zhu and Q. Ji, "Robust real-time face pose and facial expression recovery," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2006, pp. 681–688.

[16] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2562–2569.

[17] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 1202–1207.

[18] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1357–1369, Jun. 2013.

[19] F. Zhang, Q. Mao, X. Shen, Y. Zhan, and M. Dong, "Spatially coherent feature learning for pose-invariant facial expression recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1s, pp. 1–19, Mar. 2018.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[22] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[23] G. Papandreou *et al.*, "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4903–4911.

[24] J. Krause *et al.*, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Proc. ECCV*, 2016, pp. 301–320.

[25] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.

[26] 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.

[27] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR06)*, Apr. 2006, pp. 211–216.

[28] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[29] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5040–5049.

[30] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. ICLR*, 2017.

[31] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: Fine-grained image generation through asymmetric training," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2745–2754.

[32] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[33] 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.

[34] X. Tang, Z. Wang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7939–7947.

[35] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[36] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228.

[37] S. Ma, J. Fu, C. W. Chen, and T. Mei, "DA-GAN: Instance-level image translation by deep attention generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5657–5666.

[38] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[39] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[40] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[41] 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. 2242–2251.

[42] C. Ledig *et al.*, "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.

[43] 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: https://arxiv.org/abs/1710.10196

[44] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 406–416.

[45] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784

[46] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6089–6098.

[47] 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. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2867–2877.

[48] 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.

[49] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 360–368.

[50] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.

[51] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 370–376.

[52] U. Tariq, J. Yang, and T. S. Huang, "Supervised super-vector encoding for facial expression recognition," *Pattern Recognit. Lett.*, vol. 46, pp. 89–95, Sep. 2014.

[53] Q. Mao, Q. Rao, Y. Yu, and M. Dong, "Hierarchical Bayesian theme models for multipose facial expression recognition," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 861–873, Apr. 2017.

[54] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3466–3474.

[55] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.

[56] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[57] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 71–85, Jan. 2014.

[58] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 357–366.

[59] J. Gu, X. Yang, S. D. Mello, and J. Kautz, "Dynamic facial analysis: From Bayesian filtering to recurrent neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1548–1557.

[60] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[61] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," in *Affective Computing, Emotion Modelling, Synthesis and Recognition*. Rijeka, Croatia: InTech, 2008, pp. 421–440.

[62] Y.-l. Tian, T. Kanade, and J. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc.5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 2003, pp. 229–234.

[63] O. Rudovic, I. Patras, and M. Pantic, "Regression-based multi-view facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4121–4124.

[64] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. ECCV*, 2012, pp. 808–822.

[65] S. Yu, J. Wu, S. Wu, and D. Xu. *Lib Face Detection*. Accessed: 2016. [Online]. Available: https://github.com/ShiqiYu/libfacedetection/

[66] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5810–5818.

[67] 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.

[68] 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.

[69] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016, pp. 1–16.

[70] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[71] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," in *Proc. ICLR*, 2016, pp. 1–16.

[72] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive GAN for facial expression transfer," 2018, *arXiv:1802.01822*. [Online]. Available: https://arxiv.org/abs/1802.01822

[73] 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.

[74] U. Tariq, J. Yang, and T. S. Huang, "Maximum margin GMM learning for facial expression recognition," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[75] M. Jampour, T. Mauthner, and H. Bischof, "Multi-view facial expressions recognition using local linear regression of sparse codes," in *Proc. Comput. Vis. Winter Workshop Paul Wohlhart*, 2015, p 1.

[76] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467
[77] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[78] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *TISTACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–42, Feb. 2016.
[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[80] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[81] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "A novel approach to expression recognition from non-frontal face images," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1901–1908.
[82] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.
[83] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.
[84] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, p. 1.

Feifei Zhang received the Ph.D. degree from Jiangsu University, Zhenjiang, Jiangsu, China, in 2019. She is currently an Assistant Professor with Multimedia Computing Group, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include multimedia analysis, computer vision, deep learning, especially multimedia computing, facial expression recognition, and cross-modal image retrieval.

Tianzhu Zhang (Member, IEEE) received the bachelor's degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2011. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and multimedia, especially action recognition, object classification, object tracking, and social event analysis.

Qirong Mao (Member, IEEE) received the M.S. and Ph.D. degrees in computer application technology from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively. She is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University. Her research is supported by the Key Project of National Science Foundation of China (NSFC) of Jiangsu Province and the Education Department of Jiangsu Province. She has published over 50 technical articles, some of them in premium journals and conferences, such as the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE CVPR, and the ACM Multimedia. Her research interests include affective computing, pattern recognition, and multimedia analysis. She is also a member of ACM.

Changsheng Xu (Fellow, IEEE) is currently a Distinguished Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He holds 40 granted/pending patents and has published over 300 refereed research articles in these areas. He is a fellow of IAPR and a Distinguished Scientist of ACM. He has served as an Associate Editor, a Guest Editor, the General Chair, the Program Chair, the Area/Track Chair, a Special Session Organizer, the Session Chair, and a TPC Member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops, including the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications*, and the ACM Multimedia conference.