Joint Expression Synthesis and Representation Learning for Facial Expression Recognition

Xi Zhang, Feifei Zhang, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— Facial expression recognition (FER) is a challenging task due to the large appearance variations and the lack of sufficient training data. Conventional deep approaches either learn a good representation through deep models or synthesize images automatically to enlarge the training set. In this paper, we perform both tasks jointly and propose an end-to-end deep model for simultaneous facial expression recognition and facial image synthesis. The proposed model is based on Generative Adversarial Network (GAN) and enjoys several merits. First, the facial image synthesis and facial expression recognition tasks can boost their performance for each other via the unified model. Second, paired images are not required in our facial image synthesis network, which makes the proposed model much more general and flexible. Meanwhile, the generated facial images largely expand the training set and ease the overfitting problem in our FER task. Third, different expressions are encoded in a disentangled manner in a latent space, which enables us to synthesize facial images with arbitrary expressions by exchanging certain parts of their latent identity features. Quantitative and qualitative evaluations on both controlled and in-the-wild FER benchmarks (Multi-PIE, MMI, and RAF-DB) demonstrate the effectiveness of our proposed method on both facial image synthesis and facial expression recognition task.**

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

Manuscript received September 2, 2020; revised November 30, 2020 and January 5, 2021; accepted January 27, 2021. Date of publication February 1, 2021; date of current version March 9, 2022. 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 under Grant 61720106006, Grant 61721004, Grant 61832002, Grant 61532009, Grant 62002355, 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 National Postdoctoral Program for Innovative Talents under Grant BX20190367. This article was recommended by Associate Editor V. Stankovic. *(Corresponding author: Changsheng Xu.)*

Xi 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: zhangxi2019@ia.ac.cn).

Feifei Zhang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feifeizhang1231@gmail.com).

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).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3056098.

Digital Object Identifier 10.1109/TCSVT.2021.3056098

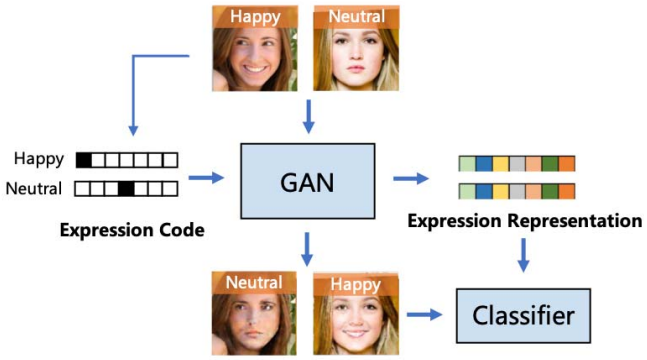


Fig. 1. Given two facial images with different expressions as input, the proposed method can produce expression representations and transfer the expressions to each other in new facial images. The learned expression representation is both generative and discriminative, which is able to improve FER performance and synthesize facial images with unpaired images. .

I. Introduction

FACIAL expression recognition (FER) is one of the most significant tasks in computer vision, which plays a crucial role in various applications such as medical care, behavior analysis, driver fatigue surveillance, and many other human-computer interaction systems [1], [2], [3]. The FER algorithm aims to classify the facial expression into several basic expressions, such as happy, sad, angry, fear, disgust, and surprise. Due to the high nonlinearity of facial expression changes and large appearance variations such as identity bias, head pose, illumination, and occlusion, the FER is a rather challenging task.

Conventional FER approaches usually conduct the facial expression recognition in the laboratory-controlled setting and in frontal or near frontal view. However, facial images are often taken from real-world environment, different angles, and appear spontaneously. Therefore, in addition to the traditional FER, many specific branches of FER warrant future study, for example, facial action unit recognition, pose-invariant FER, and FER in-the-wild. Specifically, 1) the facial action unit recognition conducts the expression recognition through detecting action units that have strong probabilistic dependencies with facial expressions from a facial image, such as lip tightening and check raising. However, the dependencies usually need further research and hard to be defined. In addition, the annotation of the action units is a cumbersome task, which usually requires professional knowledge.

2) Pose-invariant FER and FER in-the-wild tasks devote to identifying or authorizing individual facial expressions captured in arbitrary poses or real-word scenarios without any controlled conditions. Therefore, compared with the frontal or nearly frontal-based FER, the pose-invariant FER and FER in-the-wild is more challenging and more applicable. Moreover, in contrast to the facial action unit recognition, they are more convenient to obtain the training data, because only the expression labels are needed. Therefore, our paper devotes to the latter.

To conduct the pose-invariant FER and FER in-the-wild tasks with challenging factors, a wide variety of approaches have been proposed. Traditional methods often perform feature extraction via robust local descriptors such as Gabor Wavelets [4], [5], Local Binary Pattern (LBP) [6], and conduct classification using multiple Naive-Bayes classifiers [7], Support Vector Machines (SVMs) [8] or other classic classifiers. However, the representation power of the hand-crafted features is limited, and the invariance and discriminability cannot be well balanced in these methods. Therefore, the traditional methods are struggling to solve the nonlinear facial texture warping caused by large pose variation and the disturbance of various expression-unrelated factors. Recently, with the success of deep learning technology in different fields, the deep neural network has been increasingly leveraged to learn discriminative and robust representations. To tackle the nuisance factors mentioned above, large-scale labeled training data are indispensable. However, current FER datasets typically include a rather limited number of labeled facial images, which seriously prevents researchers developing excellent FER model because of the overfitting problem. To ease the overfitting problem and further improve the feature representation ability, conventional methods usually employ deep networks pre-trained on the ImageNet [9] and then do fine-tuning [10], [11], [12]. Some other methods unify several datasets together to strive for training a more general model [13], or use related information in different modalities such as landmarks and texts [14], [15] to supervise the training process. However, these methods are usually not end-to-end and rather complex. More recently, Generative Adversarial Networks (GANs) have been successfully used to generate realistic images through adversarial training. GAN-based approaches have been widely performed in face-related tasks, such as face rotation [1], [16], face attribute manipulation [17], [18], face landmark detection [19], and facial expression synthesis [20], [21]. For facial expression recognition, some researchers leverage the GAN to generate new facial images. For example, Zhang *et al.* [16] synthesize identity-preserving faces with different target poses to do the pose-invariant FER. Yang *et al.* [22] leverage a conditional GAN to generate images with neutral expressions and do the FER through de-expression residue learning. These methods can generate high-quality images to enlarge the training set and alleviate the overfitting problem, and then improve the FER results. However, they usually need paired face images with the same subject for training, which is unavailable in most existing FER datasets, particularly for the in-the-wild FER datasets.

To address the above issues, we propose an unpaired image-based generative architecture to simultaneously learn discriminative representations and synthesize facial images with arbitrary expressions, which is trained in an end-to-end manner and can achieve the state-of-the-art in both controlled and real-world FER benchmarks. As shown in Figure 1, our approach is constructed of a generative adversarial network and a facial expression classification model. Specifically, we conduct the generator in GAN with an encoder-decoder architecture to learn disentangled expression representations. The representations are not only generative to do the facial image generation, but also can be used to enhance the FER performance. It is notable that paired images with the same identity are not required to supervise the image synthesis task in our model. The inputs of the encoder G_{enc} are two unpaired facial images with different expressions, and the G_{enc} is trained to map the input to a latent space where different expressions are represented in a disentangled manner. To transfer the expressions of the two input faces to each other, a switch unit following the G_{enc} is incorporated to switch specific parts of two expression representations corresponding to the expression codes, and then the new representations are fed into G_{dec}. Through several deconvolutional layers in the G_{dec}, the outputs of the generator G are two synthesized facial images and two reconstructed facial images. In the two synthesized images, the facial expressions are exchanged successfully and the identity information is well preserved. The generator G serves as an expression transferor, and the discriminator D is trained to distinguish real images *vs.* synthesized images. Finally, an additional deep classifier C is embedded into the network, which exploits the original and synthesized facial images and the learned discriminative expression representations to do the FER task. Overall, facial image synthesis and facial expression recognition can be bridged by the latent expression representations learned by the generator, and these two parts can boost their performance for each other with the help of the generated high-quality facial images.

A preliminary version of this work was published in [23]. We extend it in numerous ways: 1) The G_{enc} in the proposed model can be used not only for image synthesis but also for FER. In other words, the learned representations from the G_{enc} are generative to produce realistic facial images, and discriminative to help Classifier C to recognize facial expressions, which is helpful in achieving more superior FER performance. 2) We add an additional interpolated loss using representation interpolation, which improves the quality of the generated images and makes the learned representations more discriminative. 3) We include considerable new experimental results such as ablation studies and quantitative facial image synthesis results (by the Fréchet Inception Distance (FID) [24]) to demonstrate the effectiveness of our proposed method on both the FER and face synthesis tasks. 4) We add another FER dataset MMI [25] in our experiment, and achieve state-of-the-art on it by the proposed method. In summary, this paper makes the following contributions.

(1) We propose an end-to-end model for simultaneous facial image synthesis and facial expression recognition. These

two tasks can boost performance for each other by jointly conducting image generation and representation learning.

(2) Paired images are not required in the GAN model when synthesizing new facial images, which makes our proposed approach more general and can be conveniently applied for the in-the-wild FER task.

(3) The proposed model is capable of synthesizing high-quality facial images with unpaired inputs, and can achieve state-of-the-art FER performance on both the laboratory-controlled FER datasets (Multi-PIE [26], MMI [25]), and in-the-wild FER dataset (RAF-DB [27]), which demonstrates the effectiveness and robustness of our approach.

II. Related Work

Facial expression recognition has attracted more and more research interests and comprehensive surveys of this field can be found in [28]. In this section, we briefly discuss the work related to the generative adversarial networks (GANs), representation learning, and facial expression recognition (FER).

Generative Adversarial Networks. Generative adversarial networks (GANs) proposed by Goodfellow *et al.* [29] have been studied vigorously and have shown impressive results in various computer vision tasks including image synthesis [16], [30]–[32], super-resolution imaging [33], [34], [35], semantic manipulation [36], [37], and so on. A typical GAN consists of a generator G and a discriminator D. The task of G is to use the random variables to synthesize fake samples, while D is trained to distinguish between the real and fake samples. Through the max-min game between G and D, G can synthesize real-like samples that we need. Concretely, the objective function of GAN can be formulated as follows:

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

where $p_d(x)$ denotes the distribution of real images x, and $p_z(z)$ is the prior distribution on the input noise variables z. The generator G and discriminator D are trained alternatively. There are many extensions of GAN: cGAN [38] is developed for the conditional generation, DCGAN [32] adopts deconvolutional and convolutional neural networks to implement G and D respectively, and cycleGAN [39] uses an encoder-decoder based generator to do image-to-image conversions. Based on these basic architectures, a variety of GANs have been proposed in face recognition and face generation. For example, Zhang *et al.* [40] propose a GAN-based model to synthesize facial images with the guidance of labeled and unlabeled data. The De-expression Residue Learning (DeRL) model [22] uses encoder-decoder based GAN to generate the corresponding neutral images for any input face. Zhang *et al.* [16] automatically synthesize facial images with arbitrary expressions under different poses with the cGAN. However, these methods all require paired images with the same subject for training, which are usually unavailable in most FER benchmarks, thus the models are not flexible. In construct to these methods, our proposed approach does not need paired images for facial image synthesis. In the proposed model, different expressions can be encoded in a disentangled manner in the latent representation, which helps to synthesize facial images with arbitrary expressions and preserve personal information. More similar to us, Xiao *et al.* [18] do the face attribute transferring (eyeglasses, gender, bangs, smile, and hair) with unpaired images. However, their method is not complemented with other relevant tasks in a joint framework. Furthermore, different from the existing studies which mainly focus on image synthesis, our proposed method leverages the discriminative and generative representations learned by the generator and obtains better results on both image synthesis and FER tasks.

Representation Learning. With the recent advances in deep learning techniques, discriminative representations can be learned from a well-trained deep neural network. The learned representations are usually insensitive to the nuisance factors in FER, such as illumination changes, pose variations, and identity bias. A lot of prior work has explored representation learning for facial expression recognition. For example, Bargal *et al.* [41] adopt different networks to extract several representations, which are then concatenated to describe the corresponding facial image. Zhang *et al.* [42] propose a multi-signal CNN (MSCNN) to conduct FER and face verification task in a unified model. Different from the existing methods, our proposed approach utilizes the generator in GAN for expression representation learning. The learned representation can be used to both facial expression classification and image synthesis, and representation interpolation technique helps these two tasks perform better. More relevant to our work, DR-GAN [1] joints face rotation and representation learning to perform pose-invariant face recognition. We are inspired by [1], but differently, each kind of expression can be encoded in a fixed position of the learned representation in our proposed model, and the learned representations can be successfully used to aid the facial expression recognition task.

Facial Expression Recognition. In recent years, facial expression recognition has attracted increasing attention from the researchers [43], [1], [22], [44]. As mentioned in [28], there are three main stages in automatic FER: pre-processing, feature learning, and feature classification. Specifically, pre-processing devotes to locating faces in complex scenes, and then aligning/normalizing the visual semantic information conveyed by the face. In feature learning, high-level abstractions are extracted from facial images through hierarchical architectures of multiple nonlinear transformations and representations. According to whether the representations are extracted by deep learning methods or by manually designed descriptors, they can be classified into learning-based representations [45], [46], [47], [48] and engineered representations [49], [50]. To overcome the overfitting problem caused by insufficient training data and large appearance variations, most methods combine different features and use some kind of unsupervised pre-training to do the model initialization. For the final stage (feature classification), the learned feature is fed into a supervised classifier (e.g., logistic regression [51], SVMs [52]) to recognize the facial expression. Different from the existing methods, the proposed method combines facial expression synthesis and facial expression recognition in an end-to-end manner. We employ the GAN to generate faces

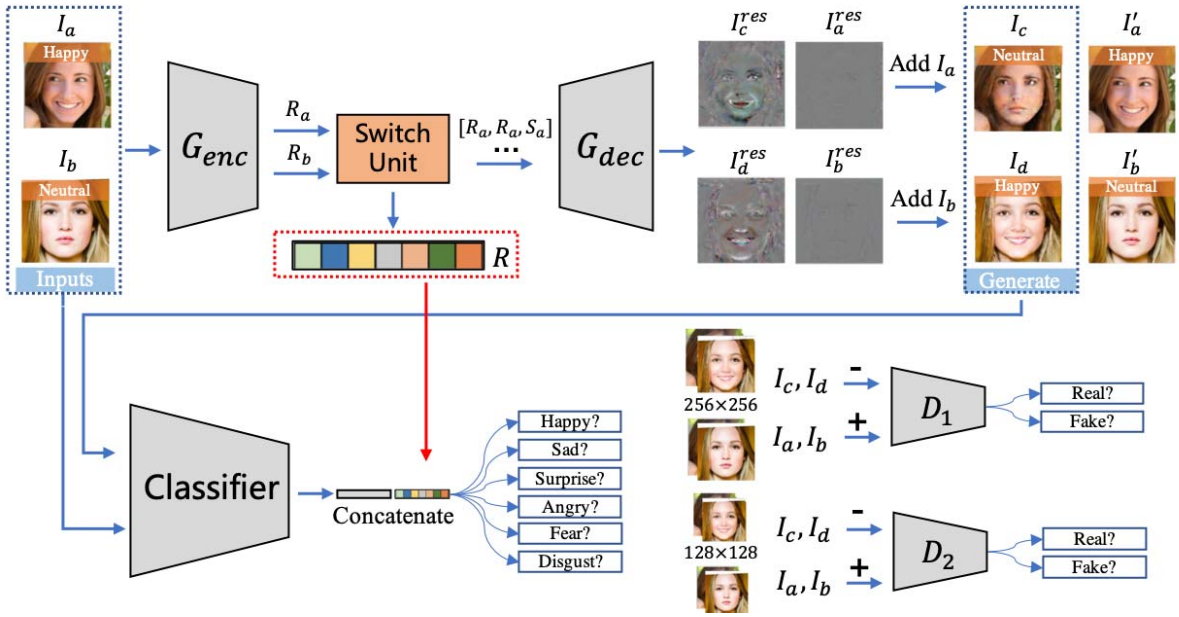


Fig. 2. The overall architecture of the proposed model, which incorporates a generator G with G_{enc}, switch unit, and G_{dec}, two discriminators D_1 and D_2, and a classifier C. With the learned generative and discriminative representation R, the proposed model can transfer expressions to generate new facial images and perform well on FER.

with arbitrary expressions to enlarge the training set, and the unpaired image setting makes the model more general for the in-the-wild facial images. Besides, the discriminative representations learned by a deep classifier and the generator can be concatenated to overcome the challenging expression-unrelated factors.

III. Proposed Method

In this section, we first introduce the architecture of the proposed network shown in Figure 2, which simultaneously conduct facial image synthesis and the representation learning for facial expression recognition. Then, we introduce the learning details of our method. Finally, we compare our model with several related models in detail.

A. Overview

We propose an end-to-end deep FER model for simultaneous facial expression recognition and unpaired facial image synthesis. The architecture of the proposed model consists of a generator G, the multi-scale discriminators D_1 and D_2, and a deep classifier C. The generator and discriminators are trained for the facial image synthesis and representation learning, while the classifier is trained for facial expression classification. In particular, we first feed two facial images with different expressions and identities into the generator G. Then the G extracts latent encodings of input facial images, exchanges certain parts of latent identity features, and generates new facial images with transferred expressions through several deconvolution layers. The multi-scale discriminators D_1 and D_2 have identical network structures but different receptive fields, which are trained to guide the generator to produce finer details and handle the holistic image content respectively. The discriminators help to make different expressions disentangled better and synthesized images more photorealistic. After the facial image synthesis, the deep classifier C is proposed to perform the FER task with the help of the original facial images, the synthesized facial images, and the expression representations learned by G.

B. Image Synthesis and Representation Learning

Generator G**.** we model the generator G as an encoder-decoder with the U-Net [10] architecture, while the switch unit is between the encoder G_{enc} and the decoder G_{dec}. Suppose n is the number of pre-defined expression categories. Given two facial images $\{I_a, I_b\}$ with different expressions (e.g., *Happy* and *Neutral* in Figure 2), their expression labels are $\{y^i, y^j\}$ $(y^i \neq y^j)$, and $1 \leq i, j \leq n$. Note that the two input images are not required to come from the same subject in our model. Our goals are: 1) to generate new facial images $\{I_c, I_d\}$ by transferring the expressions from one input image to another; 2) to learn generative and discriminative representations with different expressions disentangled.

G_{enc} is used to extract expression representations $\{R_a, R_b\}$ from the input facial images $\{I_a, I_b\}$. Notably, $\{R_a, R_b\}$ are divided into n blocks along the channel dimension. For each expression, the corresponding block in R_a (or R_b) is pre-defined. During the training, we train the generator G with respect to a particular expression each time and go over all expressions repeatedly [18]. In this way, each expression can be encoded into different and fixed part of the learned representation, and different expressions are disentangled. Specifically, the expression representations $\{R_a, R_b\}$ can be written as:

$$\begin{aligned} R_a &= G_{enc}(I_a) = [r_a^1, \ldots, r_a^i, \ldots, r_a^j, \ldots, r_a^n], \\ R_b &= G_{enc}(I_b) = [r_b^1, \ldots, r_b^i, \ldots, r_b^j, \ldots, r_b^n], \end{aligned} \tag{2}$$

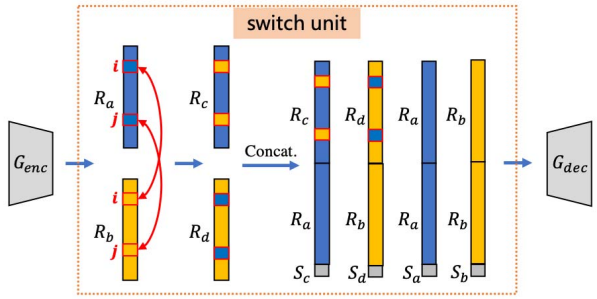


Fig. 3. The Architecture of the switch unit. In the switch unit, certain blocks of the representations which learned by the G_{enc} are exchanged according to expression labels. The switched representations are then concatenated to feed into the G_{dec} to generate residual images with transferred expressions.

where r_a^i (or r_b^i) is the representation tensor encoded the i_{th} expression (y^i) of image I_a(or I_b), and r_a^j (or r_b^j) is the representation tensor encoded the j_{th} expression (y^j).

Then the extracted representations are fed into the switch unit, in which corresponding blocks of the representation are exchanged according to expression labels. As shown in Figure 3, in the switch unit, the i_{th} and j_{th} part of R_a and R_b can be exchanged to get new latent encodings R_c and R_d,

$$\begin{aligned} R_c &= [r_a^1, \ldots, r_{\boldsymbol{b}}^i, \ldots, r_{\boldsymbol{b}}^j, \ldots, r_a^n], \\ R_d &= [r_b^1, \ldots, r_{\boldsymbol{a}}^i, \ldots, r_{\boldsymbol{a}}^j, \ldots, r_b^n]. \end{aligned} \tag{3}$$

Through the switch unit, the i_{th} and j_{th} expressions of images I_a and I_b can be transferred to each other in the representation R_c and R_d. In other words, the new generated feature R_c represents expression y^j of image I_a, and R_d is the representation of expression y^i of image I_b.

Besides, we adopt residual learning in our facial image synthesis model. To be specific, we synthesize residual image instead of a new facial image directly, which allows us to modify only a specific part of input images. Without paired faces with the same subject, the proposed model can preserve subject-related information well and alleviate the training difficulty in this way. To let G_{dec} learn the "subtraction" operation, we concatenate R_c and R_a, R_d and R_b for expression change, while concatenate R_a and R_b with themselves to synthesis reconstructed images. As shown in Figure 3, to ensure the expressions to be transferred between I_a and I_b, a binary switch code S is also concatenated into the learned representations, which represents the exchanging index. The concatenated tensors are finally fed into G_{dec} to do the facial image generation:

$$\begin{aligned} G_{dec}([R_c, R_a, S_c]) &= I_c^{res}, \quad I_c = I_c^{res} + I_a, \\ G_{dec}([R_d, R_b, S_d]) &= I_d^{res}, \quad I_d = I_d^{res} + I_b, \\ G_{dec}([R_a, R_a, S_a]) &= I_a^{res}, \quad I_a' = I_a^{res} + I_a, \\ G_{dec}([R_b, R_b, S_b]) &= I_b^{res}, \quad I_b' = I_b^{res} + I_b, \end{aligned} \tag{4}$$

where I_a^{res}, I_b^{res}, I_c^{res}, I_d^{res} represent the generated residual images, I_a' and I_b' are the reconstructed images, and I_c and I_d are the generated facial images corresponding to I_a and I_b. However, their expressions have been swapped. $[R_c, R_a, S_c]$ denotes the concatenation of representation R_c, R_a, and switch code S_c.

Discriminator D_1 and D_2. In this paper, we adopt two discriminators in our face image synthesis model. Each discriminator is constructed as a multi-scale architecture, which can operate at different image scales. Particularly, we denote D_1 as the discriminator that operating at a larger scale and D_2 as the one operating at a smaller scale, which is implemented by imposing different receptive fields on the D_1 and D_2. With this setting, the D_1 is helpful to produce finer details of the synthesized image, whereas D_2 is specialized in handling the holistic image content. Conditioned on the expression label y, the discriminator can assist G to learn disentangled representations of the facial image and then transfer expressions. Concretely, the generator and the discriminators are trained to optimize the following objective:

$$\begin{aligned} L_{adv} = \sum_{i=1}^{2} \min_G \max_{D_i} & E_{I\sim P_{im}, y\sim P_y}[log(D_i(I, y))] \\ & + E_{I\sim P_{im}, y\sim P_y}[log(1 - D_i(G(I, y), y))], \end{aligned} \tag{5}$$

where P_{im} and P_y represent the face image distribution and expression label distribution, respectively.

Representation Interpolation. To train a better generator for representation learning and image synthesis, we also employ the representation interpolation to design an interpolation loss L_{inter} to improve the learning process. For the expression y^i, we believe that the interpolation between the representation blocks r_a^i (from I_a) and r_b^i (I_b) can still generate a valid face but the intensity of expressions y_i drops, because only I_a contains information for expression y_i while I_b does not have. It is the same for the other expressions. Therefore, we can obtain the new representation R_c^p and R_d^p as follows,

$$\begin{aligned} & R_c^p \\ & = [r_a^1, \ldots, pr_a^i + (1-p)r_b^i, \ldots, pr_a^j + (1-p)r_b^j, \ldots, r_a^n], \\ & R_d^p \\ & = [r_b^1, \ldots, pr_b^i + (1-p)r_a^i, \ldots, pr_b^j + (1-p)r_a^j, \ldots, r_b^n]. \end{aligned} \tag{6}$$

We set $p = 0.5$ in the experiments for simplicity. When we input the interpolated representations R_c^p and R_d^p which have lower target expression intensities into G_{dec}, we still aim to generate the facial image with target expression y_j (e.g., *Neutral*) for image I_a, and the image with expression y_i (e.g., *Happy*) for I_b. In this way, the generator is asked to focus on target expressions and the ability of G can be further improved. Then G aims to classify these generated images to *real* class while D aims to classify them to *fake* class. The corresponding loss is named as the interpolation loss which belongs to the adversarial loss and is formulated as L_{inter} in Eq. (7):

$$\begin{aligned} l^{inter}(x) = \sum_{i=1}^{2} \min_G \max_{D_i} & E_{I\sim P_{im}, y\sim P_y}[log(D_i(G_{dec}(x)))] \\ & + E_{I\sim P_{im}, y\sim P_y}[log(1 - D_i(G_{dec}(x))], \\ L_{inter} = l^{inter}([R_c^p, R_a, S_c]) & + l^{inter}([R_d^p, R_b, S_d]). \end{aligned} \tag{7}$$

Overall, we benefit from the representation interpolation technique in two aspects. First, in our experiment, we demonstrate that G_{dec} can be improved in image synthesis with L_{inter}. With the help of the interpolation loss, our image synthesis model can generate more realistic facial images, which contributes to ease the over-fitting problem in our FER task. Second, the expression representation extracted by G_{enc} can be disentangled better in this way, which contains more discriminative and generative expression information.

C. Facial Expression Recognition

We adopt the deep model VGGNet [53] as our FER classifier. Furthermore, we also take advantage of the learned discriminative expression representation from G_{enc} for the expression classification. Specifically, the representation from G_{enc} and the deep feature extracted from the last pooling layer in VGGNet are concatenated, and then go through several fully connected layers to do the final classification. Suppose FC is the final fully connected layer and its output length equals to the number of the expression categories, we adopt a typical softmax cross-entropy loss to do the FER:

$$L_C = -E[-y^e \log FC([VGG(I), G_{enc}(I)])], \tag{8}$$

where I is the original or generated facial images and y_e represents the expression label for it. $VGG(I)$ is the deep feature extracted by the VGGNet, $G_{enc}(I)$ is the expression representation, while $[VGG(I), G_{enc}(I)]$ is the concatenated feature. With the aid of the expression representations and generated facial images, the final expression recognition results can be well improved.

D. Objective Function

In addition to the adversarial loss L_{adv} and interpolation loss L_{inter} discussed in Eq. (5) and Eq. (7), we also incorporate a construction loss and a perceptual loss to further ensure the quality of the synthesized images.

Construction Loss. To measure the reconstruction quality, we employ ℓ_2 distance to measure pixel-to-pixel differences between the reconstruction images $\{I'_a, I'_b\}$ and real images $\{I_a, I_b\}$:

$$L_r = \|I_a - I'_a\|_2 + \|I_b - I'_b\|_2. \tag{9}$$

Besides, we assume that the sum of image pixels before and after expression synthesis should be the same, because only the expressions are exchanged in the real facial images I_a and I_b:

$$L_p = \|I_a + I_b\|_1 - \|I_c + I_d\|_1, \tag{10}$$

Perceptual loss. The ℓ_1 distance is adopted for calculating the semantic difference between the original images and generated images, which is formulated as follows:

$$L_{per} = \sum_{i=1}^{N} [\|F^{(i)}(I_a) - F^{(i)}(I_d)\|_1 + \|F^{(i)}(I_b) - F^{(i)}(I_c)\|_1], \tag{11}$$

where $F^{(i)}$ denotes the i_{th} layer of the pre-trained VGG network.

Overall objective function. Finally, the overall objective function is defined as in Eq. (12) by considering the above factors:

$$\begin{aligned} L = {} & L_{adv} + L_{inter} + L_C + \lambda_1 \times L_r + \lambda_2 \times L_{per} \\ & + \lambda_3 \times L_p + \lambda_4 \times TV(G(I_a, I_b)), \end{aligned} \tag{12}$$

where λ_1, λ_2, λ_3, and λ_4 are the trade-off parameters. We also add a TV Loss during the expression synthesis, which is effective to remove the ghosting artifacts. Updating this function sequentially makes the generator extract disentangled expression representations, which is discriminative and generative, and can help our model synthesize realistic facial images with unpaired inputs. The generated facial images are able to further facilitate our FER task in both controlled and uncontrolled situations.

E. Comparison to Prior Models

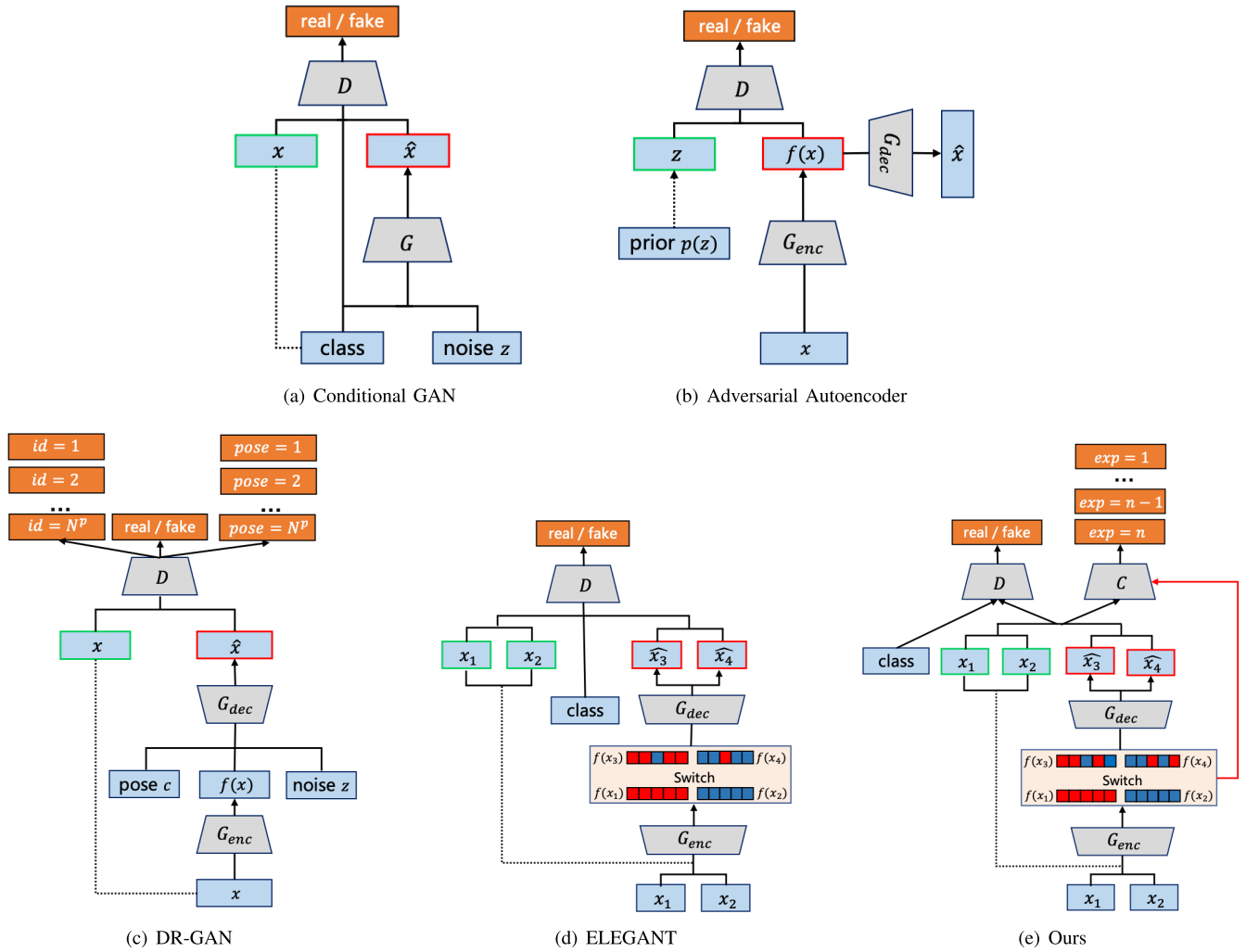
In this section, we compare the proposed model with the four most relevant approaches based on GAN as shown in Figure 4.

Conditional GAN (cGAN). The cGAN [38] is an extension of the typical GAN (Figure 4(a)), which feeds the class information to both Generator G and Discriminator D to do image generation conditioned on the labels. Different from the cGAN, the input of the generator are two facial images in the proposed method. We model the generator G with the encoder-decoder architecture to extract image representations and synthesize new facial images with transferred expressions.

Adversarial Autoencoder (AAE). As shown in Figure 4(b), AAE [54] uses the encoder of autoencoder to be the Generator G. AAE has two objectives to make the autoencoder into a generative model: the latent vector generated by G_{enc} matches an arbitrary prior distribution by training D, and the autoencoder reconstructs the input. Unlike the AAE, G_{enc} in the proposed model is trained to learn discriminative representations, in which different expressions are disentangled.

DR-GAN. In DR-GAN [1] (Figure 4(c)), the encoder-decoder in GAN learns discriminative identity representations with one or multiple inputs, while the discriminator is trained to classify real/fake images, the pose, and the identity of the input images. The DR-GAN does the pose-invariant facial recognition by performing face frontalization and learning pose-invariant representations. Similar to it, our model also does the facial expression synthesis and representation learning jointly. However, in our model, different facial expressions are trained to be disentangled in the learned expression representation, which contains rich expression information to do the synthesis and can also be used in facial expression classification straightly. Moreover, paired images with the same subject are not required in our model while the DR-GAN needs multiple images per subject in training.

ELEGANT. The ELEGANT [18] (Figure 4(d)) transfers multiple face attributes by exchanging certain parts of the encodings from two images with opposite attributes, which is most related to our proposed model. Our model differs from



(a) Conditional GAN (b) Adversarial Autoencoder

(c) DR-GAN (d) ELEGANT (e) Ours

Fig. 4. Illustration of the structure of Conditional GAN [38], Adversarial Autoencoder [54], DR-GAN [1], ELEGANT [18], and the proposed model. $x(/x_*)$ and $\hat{x}(/\hat{x_*})$ are the input and generated images, z is the noise vector, $G(/G_*)$ and D are generator and discriminator in GAN, and C is the classification network.

the ELEGANT in three aspects. First, we expand it to do the facial expression transfer, which is more difficult than the attribute generation task, because the facial expressions are combined more closely than face attributes and have higher nonlinearity. Second, we take the facial expression transfer task with the FER task together in a unified model, while the ELEGANT is only trained for face attribute transfer. To achieve this, on the one hand, we incorporate a deep classifier into our model. On the other hand, we apply the generative expression representations $f(x)$ from G_{enc} into the classifier to assist its training, and demonstrate superior FER performance even in the wild conditions. Third, during the image synthesis, an interpolated loss is adopted to make different expressions disentangled better, and further improve the image quality during the training.

IV. Experimental Results

In this section, we clarify the datasets used in our method (Sec. IV-A) and then show the implementation details of the proposed model (Sec. IV-B) including network architecture and settings. Then we show the quantitative results of our FER task. Besides, we show generated samples and evaluate the quality of them using the FID score. Finally, we conduct ablation studies to justify various design choices, including expression representations and synthesized images. The experiments are conducted on both the controlled and in-the-wild datasets.

A. Datasets

We conduct extensive quantitative and qualitative experiments on three standard FER datasets to demonstrate the effectiveness of our proposed approach. (1) Multi-PIE [26]: the controlled public multi-pose FER dataset with static images. (2) MMI [25]: the controlled public FER dataset with image sequences. (3) RAF-DB [27]: the in-the-wild FER datasets with large diversities.

Multi-PIE: The Multi-PIE dataset contains static facial images with various poses, illumination, and expression variations in the controlled setting. We follow the setting in [55] and select images of 270 subjects captured at five pan angles

30°, 15°, 0°, −15°, and −30°, which depict acted facial expressions including *Disgust, Scream, Smile, Squint, Surprise, and Neutral*. Consequently, we have 1, 521 images per pose and 7, 655 facial images in total. We perform five-fold subject independent cross-validation. As a result, the training set contains 6, 124 facial images and the test set contains 1, 531 images.

MMI: The MMI dataset contains 236 image sequences from 31 subjects in the controlled setting, and each sequence is labeled as one of the seven basic facial expressions: *Surprise, Angry, Disgust, Fearful, Happy, Surprise, Sad, and Neutral*. Following the setting in [22], we selected 208 sequences captured in the frontal view. For each sequence, we select the first frame as the *Neutral* expression image while select three frames near the middle as peak frames. These peak frames are associated with the given expression labels. Finally, we get a dataset containing 830 images. Besides, we perform five-fold subject independent cross-validation. As a result, the training set contains 656 images, while the test set contains 174 images.

RAF-DB: The RAF-DB is a large-scale facial expression dataset with 30K great-diverse real-world facial images downloaded from the Internet. The images in this dataset are of the great variability in subjects' head poses, age, ethnicity, gender, ethnicity, occlusions, lighting conditions, and post-processing operations, which is challenging for the FER task. We only use the images with basic emotions (*Sad, Surprised, Disgusted, Happy, Fearful, Angry, and Neutral*) in our experiment. Following the official partition in [27], the training set contains 12,271 images while the test set contains 3,068 images. For all the three datasets, the training set and test set are partitioned based on the identity, thus the subjects in the two subsets are mutually exclusive.

B. Network Architecture and Settings

We construct our network according to Figure 2. Before feeding the facial images into the proposed model, we crop out the faces with 68 facial landmarks using a lib face detection algorithm [56] and resize them as 256×256. The intensities of input images are normalized into the range $[-1, 1]$ to make the training process converge faster. As shown in Figure 2, the proposed model consists of three parts: a generator G, a multi-scale discriminator $\{D_1, D_2\}$, and a deep classifier C. The detailed network structure of our model is shown in Table I. To stabilize the training process, the GAN architecture is based on the techniques in the ELEGANT [18]. In detail, G_{enc} is composed of 5 convolution (Conv) layers and G_{dec} has 5 deconvolution (DeConv) layers, while each layer is followed by l_2-normalization (l_2-N) [18] and LeakyReLU (LReLU) layers. The output of the G_{enc} is the expression representation R, which is then fed into the switch unit to do the expression transfer and concatenation. Through several deconvolutional layers, G_{dec} transforms the concatenated representations into residual images, which have the same shape as input images. The multi-scale discriminator is equipped with 4 blocks (Conv, l_2-N, LReLU) followed by a fully connected (FC) layer, which is learned to identify whether the input images are real or fake. The size of images fed into D_1 is 256×256 while the size of images fed into D_2 is 128×128. For the Classifier C, we adopt the VGGNet-19 as the deep feature extractor. During the test, we feed the test image into the trained G_{enc} to get the expression representation R, which is then concatenated with the deep feature before the last pooling layer (AvgPool), and then do the final classification.

TABLE I

THE STRUCTURE OF THE PROPOSED METHOD. IN $((De)Conv(d, k, s))$, THE PARAMETER d, k AND s STAND FOR THE NUMBER OF FILTERS, THE KERNEL SIZE, AND THE STRIDE. n IS THE NUMBER OF PRE-DEFINED EXPRESSION CATEGORIES

Generator G	**Discriminator** D_1, D_2
Conv(64, 3, 2), l_2-N, LReLU	Conv(64, 3, 2), l_2-N, LReLU
Conv(128, 3, 2), l_2-N, LReLU	Conv(128, 3, 2), l_2-N, LReLU
Conv(256, 3, 2), l_2-N, LReLU	Conv(256, 3, 2), l_2-N, LReLU
Conv(512, 3, 2), l_2-N, LReLU	Conv(512, 3, 2), l_2-N, LReLU
Conv(512, 3, 2), l_2-N, LReLU	FC(1)
Concat(512 + 512 + 20 × n)	**Classifier** C
DeConv(512, 3, 2), l_2-N, LReLU	VGGNet-19(512)
DeConv(256, 3, 2), l_2-N, LReLU	Concat(512 + 512)
DeConv(128, 3, 2), l_2-N, LReLU	FC(512)
DeConv(64, 3, 2), l_2-N, LReLU	AvgPool(7, 7)
DeConv(3, 3, 2), Tanh	FC(4096), ReLU
	FC(4096), ReLU
	FC(n)

We implement our proposed method in PyTorch. Specifically, ADAM with learning rate 0.0002, $b_1 = 0.5$ and $b_2 = 0.999$ is adopted to optimize the GAN. The VGGNet is pre-trained on ImageNet [9], and optimized with SGD that initialized with learning rate 0.001 and momentum 0.9. The trade-off parameters in Eq. 12 are set to be $\lambda_1 = 10, \lambda_2 = 15, \lambda_3 = 5, \lambda_4 = 5$. The total number of parameters of our end-to-end model is 154.1M. In details, the VGGNet-based facial expression recognition module has 140.1M parameters, while the GAN-based facial expression synthesis module has 14M parameters which is much smaller than the GAN. Besides, based on a platform with a TITAN Xp GPU and the IntelCore CPU i7-6700K (4.0GHz), on each dataset, the image synthesize time is 12.98ms per image, and the facial expression recognition time is 12.06ms per image.

C. Results of Facial Expression Recognition

1) ***Experiments on the Multi-PIE Dataset****:* We evaluate the pose-invariant FER performance of the proposed model on the Multi-PIE dataset with kNN, LDA, LPP, GMLPP, MvDA, DS-GPLVM reported in [55], and deep learning methods including VGGNet-19 [53], PhaNet [57], and CG-FER [16]. The VGGNet-19 is pre-trained on ImageNet and then fine-tuned on the same training set as our model. PhaNet and CG-FER are designed specifically for the pose-invariant FER. In detail, PhaNet proposes a pose-adaptive hierarchical attention network to recognize expressions and poses together. CG-FER generates facial images with arbitrary expressions and poses to enhance the performance of their FER model.

Table II summarizes the detailed results across all poses and reports the mean FER accuracy in the last column. We can find that our model achieves 93.66% recognition rate on

TABLE II

COMPARISON RESULTS OF THE FER ACCURACY FOR EACH POSE ON THE MULTI-PIE DATASET. THE HIGHEST ACCURACY FOR EACH POSE IS HIGHLIGHTED IN BOLD

Methods	Poses					Average (%)
	-30	-15	0	15	30	
kNN [55]	80.88	81.74	68.36	75.03	74.78	76.15
LDA [55]	92.52	94.37	77.21	87.07	87.47	87.72
LPP [55]	92.42	94.56	77.33	87.06	87.68	87.81
GMLPP [55]	91.86	94.13	78.16	87.22	87.36	87.74
MvDA [55]	92.49	94.22	77.51	87.10	87.89	87.84
DS-GPLVM [55]	93.55	**96.96**	82.42	89.97	90.11	90.60
VGG19 [53]	89.93	90.91	89.58	91.47	91.75	90.72
CG-FER [16]	90.76	94.72	89.11	**93.09**	91.30	91.08
PhaNet [57]	–	–	–	–	–	93.53
Ours	**93.83**	95.13	**93.16**	92.78	**93.39**	**93.66**

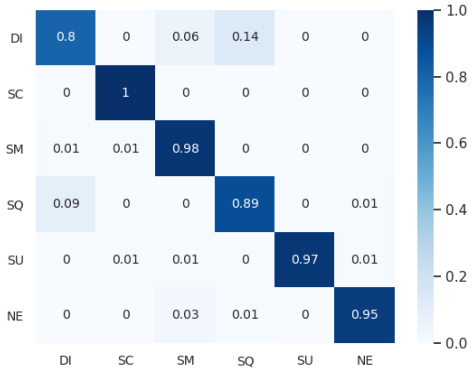


Fig. 5. Confusion matrix on the Multi-PIE dataset. The average recognition rate is 93.66%.

average and exceeds others with a satisfactory improvement from 0.13% to 17.51%. Besides, although most algorithms are not able to achieve good performance in the frontal view, the proposed method can improve it successfully and exceed the others with the growth from 3.58% to 24.8%. In Table II, we observe that all models achieve asymmetric performances and most of them achieve higher recognition accuracy for the poses $-30°$ and $-15°$. We believe the asymmetric performance is caused by the data but not the models. To conclude, the proposed approach achieves the most competitive results on the multi-pose FER dataset Multi-PIE, which benefits from the generated facial images and the discriminative representations learned by our GAN model. The confusion matrix on the Multi-PIE dataset is shown in Figure 5. From Figure 5, we can observe that the *disgust* (DI) is relatively hard to recognize, which is mostly confused with *squint* (SQ). Among all the expressions, the *scream* (SC) gets the highest recognition accuracy, which may be because the *scream* has relatively large muscle deformation than the other expressions.

*2) **Experiments on the MMI Dataset**:* In this section, we evaluate our method on the MMI dataset. Table III reports the average accuracy of basic expression recognition. Because the MMI dataset consists of image sequences rather than static images, we do comparisons with two kinds of methods. First, the sequence-based methods such as LBP-TOP [58], HOG-3D [59], F-Bases [60], and STM-ExpLet [61] employ the temporal information extracted from the video sequences for FER. Second, the image-based methods such as DeRL [22], IACNN [62], APM-VGG [63], and IF-GAN [64] only use

TABLE III

COMPARISON OF THE AVERAGE RECOGNITION ACCURACY WITH THE SEQUENCE-BASED AND IMAGE-BASED METHODS ON MMI DATASET

Methods	Setting	Classes	Accuracy (%)
LBP-TOP [58]	sequence-based	6	59.51
HOG-3D [59]	sequence-based	6	60.89
F-Bases [60]	sequence-based	6	73.66
STM-ExpLet [61]	sequence-based	6	75.12
DeRL [22]	image-based	6	73.23
IACNN [62]	image-based	6	71.55
APM-VGG [63]	image-based	6	74.04
IF-GAN [64]	image-based	6	74.52
Ours	image-based	6	**79.23**
	image-based	7	**76.44**

static images to do facial expression recognition. LBP-TOP and HOG-3D are based on hand-crafted features while other methods are based on deep learning features. F-Bases represents facial expression variations as a linear combination of localized basis functions. STM-ExpLet conducts temporal alignment and learns semantics-aware dynamic representation for FER. IACNN employs the identity-sensitive contractive loss to learn identity-related information, APM-VGG proposes adaptive pooling maps for CNNs, and IF-GAN uses an identity-free GAN to reduce inter-subject variations.

From Table III, we can observe that our method obtains 79.23% recognition accuracy on 6 expressions, and 76.44% on 7 expressions (plus *neutral* expression). Specifically, for the expression recognition on 6 expressions, our model outperforms the sequence-based methods by 4.11% to 19.72%, and exceeds the image-based methods by 4.71% to 7.68%. Furthermore, Figure 6 shows the confusion matrix on MMI datasets, which reports the average accuracies for each expression. We can find that *fear* (FE) is relatively difficult to identify, which is mostly confused with *sad* (SA) and *surprise* (SU). There are 38% of *fear* samples misclassified to be *surprise* and 25% to be *sad*. On the other hand, *surprise* (SU) can be accurately recognized with the 100% accuracy rate. The high confusion in *fear* may be caused by the insufficient training data and the similar deformations between these similar expressions. In conclusion, the proposed model can do really well in the controlled setting. The deep representations learned from original and generated images are good at employing the expression-related information, and are comparable to the dynamic representations extracted from facial image sequences.

*3) **Experiments on the RAF-DB Dataset**:* To demonstrate the robustness and effectiveness of our proposed model, the experiments on a more challenging dataset RAF-DB is also conducted. In RAF-DB, images are captured in an uncontrolled environment and have large variations in expression-unrelated factors such as head pose, background, illumination, and identity. However, as shown in Table IV, we can still obtain pretty good results. We compare our model with baseline methods AlexNet, VGG, baseDCNN, and DLP-CNN [27], and the state-of-the-art including PAT-CNN [65], APM-VGG [63], ACNN [66], and CP-CNN [67]. For the baseline models, the deep features are extracted from its penultimate FC layer and then fed into a SVM to do the expression classification. For the state-of-the-art, PAT-CNN learns

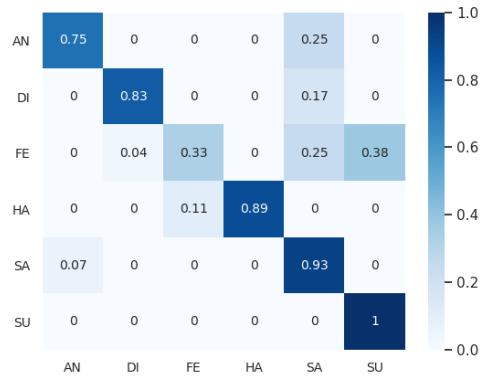


Fig. 6. Confusion matrix of 7 expressions recognition on MMI dataset. The average recognition rate is 76.44%.

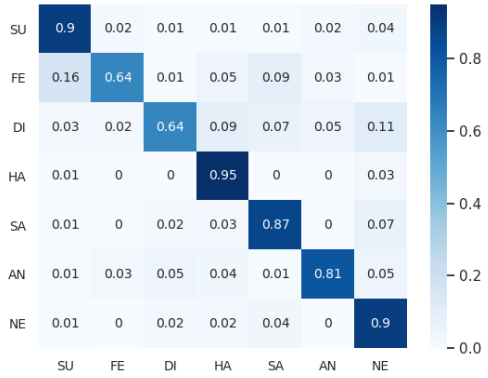


Fig. 7. Confusion matrix on RAF-DB. The average recognition rate is 89.01%.

TABLE IV

COMPARISON OF THE AVERAGE ACCURACY FOR FER WITH BASELINE METHODS AND STATE-OF-THE-ART METHODS ON RAF-DB

Methods	Accuracy (%)
VGG	58.22
AlexNet	55.60
baseDCNN	72.42
DLP-CNN [27]	74.2
PAT-CNN [65]	84.19
APM-VGG [63]	85.17
ACNN [66]	85.07
CP-CNN [67]	87.00
Ours	**89.01**

features in a hierarchical tree structure organized according to attributes. APM-VGG proposes adaptive pooling maps (APMs) for CNNs. ACNN perceives the occlusion regions of faces and focuses on the most discriminative un-occluded regions. CP-CNN adopts covariance pooling to capture the regional distortions of facial images and improves FER results.

The average FER accuracies across all facial expressions are shown in Table IV, and the confusion matrix is shown in Figure 7. From Table IV we can find that the proposed model achieves over 89% recognition rate, outperforming the compared state-of-art methods from 2.01% to 4.82%, and offering a compelling error reduction for the baseline models from 14.81% to 33.41%. In detail, from Figure 7, we can observe that *disgust* (DI) and *fear* (FE) have lower recognition accuracy while *happy* (HA) is relatively easy to identity, which probably due to that *disgust* and *fear* have smaller muscle deformation compared with *happy*. In conclusion, the experiment results on RAF-DB demonstrate that the proposed method is suitable for the in-the-wild FER. It may benefit from our unpaired facial image synthesis, which makes the model more general and robust, and helps us ease the overfitting problem in real-world FER task successfully.

D. Results of Image Synthesis

The proposed model can do facial image synthesis without paired images of the same subject. To evaluate our image synthesis model, we show the generated images qualitatively and report the FID score qualitatively in both controlled and uncontrolled settings.

*1) **Image synthesis in the controlled settings**:* Figure 8 shows the qualitative results on the Multi-PIE dataset, and the

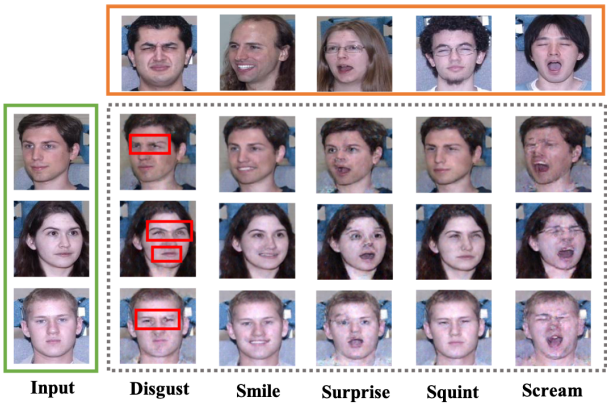


Fig. 8. Example results of the image synthesis in the controlled settings. The green and yellow rectangle represent the two sets of the input images. The gray dashed rectangle represents generated images. For each row of the generated samples, they are expected to have different expressions which transferred from the images in yellow rectangle, but the same personal identity with the image in green rectangle.

generated images are indicated in the gray dashed rectangle. The yellow rectangle and green rectangle are two sets of input images with different expressions and personal identity. We aim to synthesize new images for the facials shown in the green rectangle by transferring the expressions from the yellow rectangle. From Figure 8, we can observe that although the input images are with different identities, poses, and genders, our generative model can successfully transfer the expression and preserve the subject-related information. It implies that our approach is capable to disentangle different expressions in the expression representation, which is generative to synthesize facial images with unpaired inputs. In particular, the pose-relevant information is well preserved before and after the expression transferring, which benefits to the pose-invariant FER. Unfortunately, we can also find some distortion and blur in the generated images of *scream* expression, which may due to the larger muscle deformation. The red rectangles show the expression changes between the original images and the generated images clearly, which mostly happen to the eyes, eyebrows, and mouths.

To quantitatively evaluate the quality of the synthesized facial images, we calculate the Fréchet Inception Distance (FID) [24], which measures the distance between two

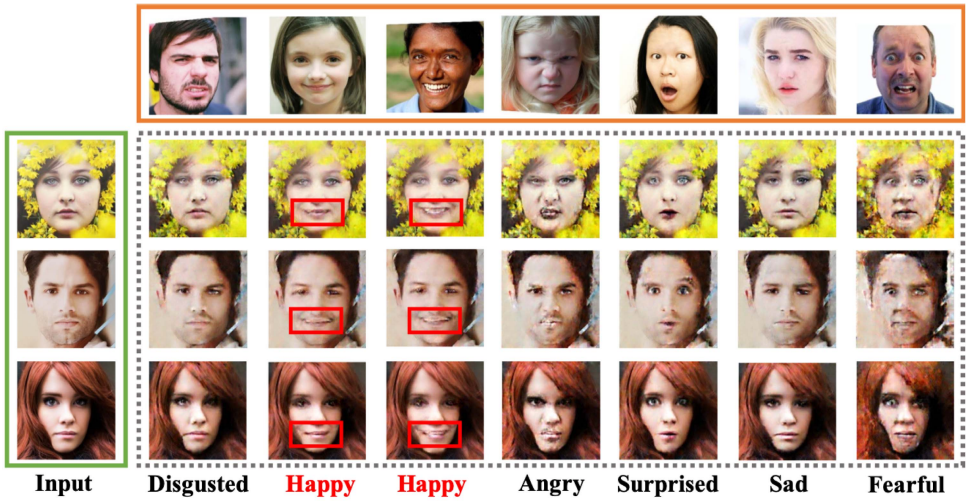


Fig. 9. Image synthesis results in the uncontrolled setting. Generated images with different transferred expressions are shown in the gray dashed rectangle. Two sets of input images are shown in the yellow and green rectangle. According to input images, the different styles of *happy* can be generated precisely (red rectangles).

distributions by:

$$d^2 = \|m - m_w\|_2^2 + \mathrm{Tr}(C + C_w - 2(CC_w)^{1/2}) \qquad (13)$$

where (m, C) and (m_w, C_w) are the means and covariance matrices of two distributions, respectively. In our experiment, the FID is computed between the distribution of the generated images and the original images with the same expression, and the results are shown in Table V. In detail, **No_inter** omits the interpolation loss when training the GAN part. By comparison, we find that the FID decreases obviously for all expressions in **Ours**, which indicates that the image quality improves successfully with the representation interpolation technique, and it is essential to train GAN with the additional interpolated expression representation.

*2) **Image synthesis in the uncontrolled settings**:* Thanks to the unpaired images based generative architecture, we can also do the facial expression synthesis vigorously in the real world, where the images are of great variabilities in lighting conditions, occlusions, background, and personal identity. We show the real-world image synthesis results on RAF-DB in Figure 9. The facial images shown in the yellow rectangle and green rectangle are the input images, while the images shown in the grayed dashed rectangle are the generated images. We aim to synthesize facial images which have the expressions transferred from the images in the yellow rectangle. The results indicate that although the facial expression synthesis model is trained with unpaired images, the proposed model is robust to the irrelevant expressions factors and can synthesize photorealistic images, which is helpful for the in-the-wild FER task and makes our model more general. Furthermore, the 3rd and 4th columns in Figure 9 would be a good illustration of generating different styles of *happy*. The expression shown in the 3rd column is the *gentle smile* without teeth, while the expression in the 4th column is the *toothy smile*. The facts indicate that the style of *happy* and other expressions from the input images can be transferred to the generated images precisely, which is benefit from the disentangled representations and the switch unit. Unfortunately, we also observe some artifacts appear in the synthesized images for *fearful* and *angry*, which may due to the insufficient training data and more complicated muscle deformation. To make the proposed method more convincing, in Figure 10, we show additional image synthesis results that contain challenging factors, such as large pose variations and occlusions.

TABLE V

THE FID RESULTS OF THE PROPOSED MODEL (**OURS**) AND THE MODEL OMITS INTERPOLATION LOSS (**NO_INTER**) ON THE MULTI-PIE DATASET. LOWER FID VALUES MEAN BETTER IMAGE QUALITY AND DIVERSITY

FID	Disgust	Scream	Smile	Squint	Surprise
No_inter	61.19	106.16	35.59	53.34	93.49
Ours	**58.23**	**99.38**	**35.15**	**52.48**	**86.16**

TABLE VI

THE FID RESULTS OF THE PROPOSED MODEL (**OURS**) AND THE MODEL OMITS INTERPOLATION LOSS (**NO_INTER**) ON RAF-DB. LOWER FID VALUES MEAN BETTER IMAGE QUALITY AND DIVERSITY

FID	Surprised	Fearful	Disgusted	Happy	Sad	Angry
No_inter	77.14	194.04	85.52	49.88	67.60	104.41
Ours	**76.30**	**183.90**	**77.93**	**40.28**	**62.95**	**88.13**

We also compute the FID between the original and generated images on RAF-DB, and the results are shown in Table VI. The **No_inter** is the model without interpolation loss. Overall, for the challenges and variability in the uncontrolled setting, the FID results on RAF-DB are mostly larger than those on the Multi-PIE dataset except for *surprise*, which may be due to the fact that the RAF-DB contains more facial images with *surprise* expression, and thus the *surprise* is well trained than the other expressions. Besides, from Table VI, we can also observe that our model still has lower FID scores and offers a compelling FID reduction compared with **No_inter**, especially for the expression *fearful* and *angry*. The results further show the effectiveness of the representation interpolation in our approach.

*3) **Representation interpolation results**:* To further evaluate whether the representation interpolation method helps to improve the quality of the generated images, we qualitatively show the discrepancy of the two models **With_inter** (With

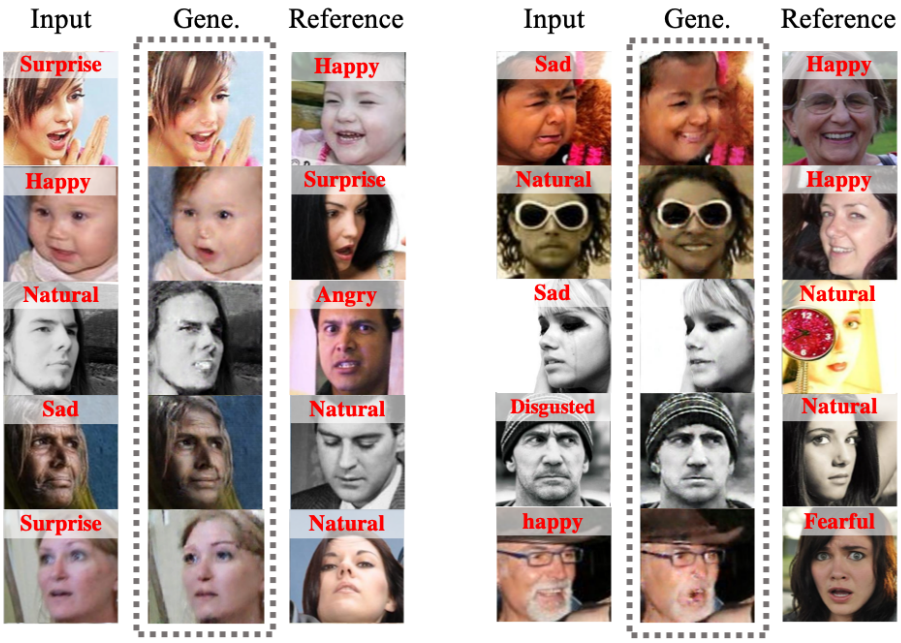


Fig. 10. Additional image synthesis results in the uncontrolled setting. The generated images are shown in the gray dashed rectangle, which are expected to have the expression of the reference image and preserve the personal identity of the input image. Specifically, the input images and the reference images contain challenging factors, such as large pose variations and occlusions (e.g., glasses, hat, and accessories).

the interpolated loss) and **No_inter** (Without the interpolated loss). In Figure 11, $\{I_1, I_2\}$ are the two input images with different expressions, the column **No_inter** shows the images which are generated by the model without the interpolated loss while the column **With_inter** shows the images which are generated with the loss. The expression of I_2 is expected to be transferred to the facials of I_1 in the synthesized images. Comparing the two columns in Figure 11, we can find that the synthesized images in the column **With_inter** have fewer blurred areas and clearer facial expressions, and facial expressions are more similar to the expression of image I_2. Overall, the result demonstrates that the interpolation loss leads to better performance in image synthesis, and the proposed model can synthesize high-quality facial images even in the wild.

E. Ablation Studies

*1) **Model Analysis**:* To show the efficacy of each module in the proposed approach, we develop several ablated versions of our method. Specifically, six variants of our method are trained on the Multi-PIE dataset.

(a). **No Interpolation Loss (No_InterL)** omits the interpolation loss L_{inter} in Eq. (12), thus the images generated from the interpolated representations are not used to train the image synthesis module.

(b). **No expression representation (No_ExpR)** omits the expression representations $G_{enc}(I)$ extracted by G_{enc} in Eq. (8) when training the FER classifier. It means that only the features extracted by the VGGNet are used to do facial expression recognition.

(c). **No Interpolation Loss and expression representation (No_InterL_ExpR)** omits the interpolation loss L_{inter} in Eq. (12) and the expression representations $G_{enc}(I)$ in Eq. (8). To be specific, the images generated from the interpolated representations are not used to train the GAN, and only the features learned by the VGGNet are adopted to train the FER model.

(d). **No synthesized facial images (No_SynFI)** only uses original facial images to train the classifier C, and the expression representation $G_{enc}(I)$ in Eq. 8 is excluded during the training.

(e). **No end-to-end and adopt ResNet50 (TwoStage_res50)** omits the end-to-end strategy used in our model, hence the facial image synthesis part and facial expression recognition part are trained separately. Besides, we replace the VGGNet-19 with ResNet50.

(f). **No end-to-end (TwoStage)** only omits the end-to-end strategy used in our model.

Table VII shows the performance of the six settings above. The fact indicates that the result from our proposed model (**Ours**) is better than all other variants and has significant improvements in almost all kinds of poses. The FER performance drops 0.59% in recognition accuracy without the interpolation loss (**No_InterL**), which shows the effectiveness of the representation interpolation in the GAN training. Without expression representation $G_{enc}(I)$ (**No_ExpR**), the FER accuracy drops 0.92%, which shows the importance of the learned discriminative representation for facial expression classification. With the representation as the bridge, the facial image synthesis and FER become co-adapt and can boost their performance for each other. Without interpolation loss and expression representation (**No_InterL_ExpR**), the accuracy drops 1.63%, which intuitively shows the effectiveness of these two parts. Furthermore, we find that the accuracy drop is larger than the sum of the drops in **No_InterL** and **No_ExpR**. This observation naturally evidences that the interpolation loss makes the representation extracted by G_{enc} be disentangled

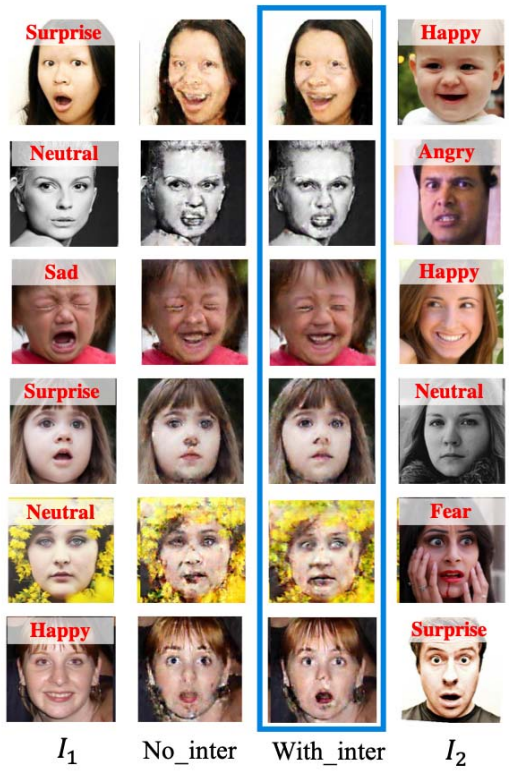


Fig. 11. Comparison of the representation interpolation results between the model trained with interpolation loss (**With_inter**) and the model without the loss (**No_inter**). $\{I_1, I_2\}$ are two input images with different expressions. The 2nd column and the 3rd column are the synthesized images with the identity of I_1 and the expression of I_2. The generated images in column **With_inter** have fewer blurred areas and clearer facial expressions.

TABLE VII

EFFECTS OF DIFFERENT COMPONENTS AND STRATEGIES ON FACIAL EXPRESSION CLASSIFICATION ACCURACY (%)

Methods	Poses					Average
	-30	-15	0	15	30	
No_InterL	92.78	93.06	92.83	93.50	93.18	93.07
No_ExpR	91.47	**94.05**	91.85	93.18	93.18	92.74
No_InterL_ExpR	91.80	92.40	92.83	89.93	93.18	92.03
No_SynFI	89.93	90.91	89.58	91.47	91.75	90.72
TwoStage_res50	90.58	93.83	90.23	90.49	91.75	91.38
TwoStage	91.23	92.86	89.58	92.46	91.09	91.44
GGPI_TwoStage [68]	90.37	90.97	87.54	87.00	92.14	89.55
StarGAN2_TwoStage [69]	91.56	91.56	91.21	92.46	**93.73**	92.09
Ours	**92.78**	93.39	**93.16**	**93.83**	**95.13**	**93.66**

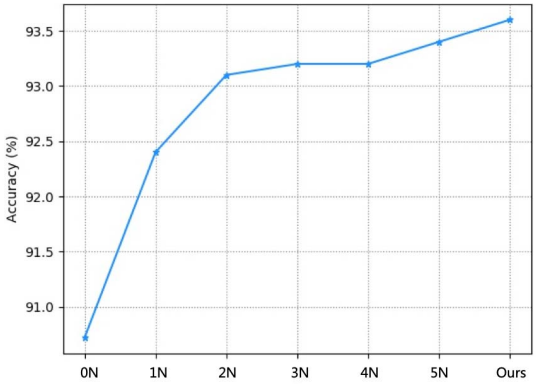


Fig. 12. Effects of the different training samples generated by the proposed model for facial expression recognition.

better and more discriminative. In particular, when we omit synthesized images (**No_SynFI**), the accuracy decreases dramatically for all poses and the average accuracy drops 2.94%. The result indicates the importance and effectiveness of our high-quality synthesized images for deep facial expression recognition.

To further show the efficacy of the facial expression synthesis module and the end-to-end strategy, we develop **TwoStage_res50** and **TwoStage** and do comparisons with **GGPI_TwoStage** [68] and **StarGAN2_TwoStage** [69]. **GGPI_TwoStage** and **StarGAN2_TwoStage** both adopt state-of-the-art GAN networks as the expression synthesis model. Specifically, **GGPI_TwoStage** first trains the GGPI [68] to generate face images with different expressions and poses by exploiting the face geometry information, then feeds the generated images into the ResNet50 to do the pose-invariant FER. It is a two-stage variant of the original framework, which is evaluated in [68]. For **StarGAN2_TwoStage**, we first train StarGAN v2 [69] to generate facial images with different expressions following the settings in [69], and then feed the generated images into VGGNet19 to do FER. From table VII, we can find that **TwoStage_res50** outperforms **GGPI_TwoStage** by a large improvement (1.82%) on the average recognition accuracy, and exceeds it for almost all poses. The fact demonstrates the superiority of our facial expression synthesis module. We also observe that **StarGAN2_TwoStage** outperforms **TwoStage** by a slight improvement (0.65%) on the FER accuracy. However, StarGAN v2 employs multi-task learning strategy, which designs multiple output branches for different expressions in mapping network, style encoder, and discriminator [69]. Therefore, it has much larger model size and needs longer training time. Differently, through a single generative network, our proposed approach can generate facial images with different expressions efficiently. Specifically, on a single Tesla V100 GPU, the training time of StarGAN v2 is about **65** hours while the training time of our facial expression synthesis model is only about **22** hours. Therefore, considering the accuracy and efficiency comprehensively, the proposed model is comparable with the state-of-the-art GANs such as StarGAN v2. Furthermore, comparing **Ours** with the two-stage frameworks, we can find that our proposed method exceeds **TwoStage** by 2.22% and exceeds **StarGAN2_TwoStage** by 1.67% on the average FER accuracy, which benefits from joint modeling of image generation and representation learning and highlights the importance of our end-to-end training strategy.

*2) **Effects of the generated facial images on FER**:* To evaluate whether the generated facial images work well in facial expression recognition, we compare our method with several models trained with the different number of synthesized images. Given the original N images, our model is trained with $6 \times N$ generated images. Besides, we randomly select $0 \times N$, $1 \times N$, $2 \times N$, $3 \times N$, $4 \times N$, and $5 \times N$ images from the generated images. The synthesized images are incorporated with the original images to train these models, where $0 \times N$ means that only the original images are used to train the classifier. Specifically, we denote them as $0N$, $1N$, $2N$, $3N$, $4N$, $5N$ to train our FER model. The average classification

rate of our method with different number of training samples is shown in Figure 12. From Figure 12, we can observe that the FER accuracy can be improved with the increase of the synthesized images. Comparing to the model trained with original images ($0N$), we can get a compelling improvement even with $1N$ generated images added. These results verify the importance of generating more training images in FER.

V. Conclusion

In this paper, we propose a deep end-to-end facial expression recognition model for simultaneous expression transfer and representation learning. Through disentangling different expressions from the facial images and representation interpolation, we can obtain both generative and discriminative representations. The learned representation is not only used to generate more training samples with unpaired input images, but also benefits to our FER tasks and helps to achieve better FER performance. The experiments on three standard datasets with controlled and in the wild settings indicate the robustness and effectiveness of our proposed approach. In future, we will attempt to further improve the quality of the generated images and do the FER task in more challenging settings.

References

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

[2] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6897–6906.

[3] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[4] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.

[5] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recognit.*, vol. 43, no. 3, pp. 972–986, Mar. 2010.

[6] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[7] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 160–187, Jul. 2003.

[8] Y. Luo, C.-M. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of PCA and LBP with SVM," *Optik, Int. J. Light Electron Opt.*, vol. 124, no. 17, pp. 2767–2770, Sep. 2013.

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

[10] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," 2017, *arXiv:1711.04598*. [Online]. Available: http://arxiv.org/abs/1711.04598

[11] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2013, pp. 543–550.

[12] T. Kaneko, K. Hiramatsu, and K. Kashino, "Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks," in *Proc. 24th ACM Int. Conf. Multimedia (MM)*, Oct. 2016, pp. 327–331.

[13] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 443–449.

[14] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Proc. Can. Conf. Comput. Robot Vis.*, May 2014, pp. 98–103.

[15] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," 2018, *arXiv:1802.06664*. [Online]. Available: http://arxiv.org/abs/1802.06664

[16] 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.

[17] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 417–432.

[18] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with GAN for transferring multiple face attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 168–184.

[19] H. J. Lee, S. T. Kim, H. Lee, and Y. M. Ro, "Lightweight and effective facial landmark detection using adversarial learning with face geometric map generative network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 771–780, Mar. 2020.

[20] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proc. 26th ACM Int. Conf. Multimedia (MM)*, Oct. 2018, pp. 627–635.

[21] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 15, 2020, doi: 10.1109/TCSVT.2021.3056098.

[22] 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.

[23] X. Zhang, F. Zhang, and C. Xu, "Unpaired images based generator architecture for facial expression recognition," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

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

[25] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, p. 5.

[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] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.

[28] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TCSVT.2021.3056098.

[29] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2016, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[30] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, p. 4570.

[31] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5077–5086.

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

[33] 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.

[34] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 136–144.

[35] P. W. Kim, "Image super-resolution model using an improved deep learning-based facial expression analysis," *Multimedia Syst.*, pp. 1–11, 2020, doi:10.1007/s00530-020-00705-1.

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

[37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

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

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

[40] F. Zhang, T. Zhang, Q. Mao, L. Duan, and C. Xu, "Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach," in *Proc. 26th ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2018, pp. 126–135.

[41] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Oct. 2016, pp. 433–436.

[42] 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.

[43] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "A unified deep model for joint facial expression recognition, face synthesis, and face alignment," *IEEE Trans. Image Process.*, vol. 29, pp. 6574–6589, 2020.

[44] X. Xiang and T. D. Tran, "Linear disentangled representation learning for facial actions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 12, pp. 3539–3544, Dec. 2018.

[45] B. Sun *et al.*, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 497–502.

[46] J. Li *et al.*, "Facial expression recognition with faster R-CNN," *Procedia Comput. Sci.*, vol. 107, pp. 135–140, Jan. 2017.

[47] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Oct. 2016, pp. 445–450.

[48] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," *Multimedia Syst.*, vol. 26, no. 5, pp. 535–551, Oct. 2020.

[49] 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. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 357–366.

[50] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[51] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 6, 2012, pp. 808–822.

[52] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[54] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*. [Online]. Available: http://arxiv.org/abs/1511.05644

[55] 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.

[56] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[57] Y. Liu, J. Peng, J. Zeng, and S. Shan, "Pose-adaptive hierarchical attention network for facial expression recognition," 2019, *arXiv:1905.10059*. [Online]. Available: http://arxiv.org/abs/1905.10059

[58] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[59] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[60] 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.

[61] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1749–1756.

[62] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.

[63] Z. Li *et al.*, "Pooling map adaptation in convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1108–1113.

[64] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," 2019, *arXiv:1903.08051*. [Online]. Available: http://arxiv.org/abs/1903.08051

[65] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," 2018, *arXiv:1812.07067*. [Online]. Available: http://arxiv.org/abs/1812.07067

[66] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[67] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 367–374.

[68] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4445–4460, 2020.

[69] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.

Xi Zhang received the bachelor's degree in computer science from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Multimedia Computing Group, Institute of Automation, Chinese Academy of Sciences, Beijing. Her research interests include multimedia analysis, computer vision, deep learning, especially multimedia computing, facial expression recognition, and visual reasoning.

Feifei Zhang received the Ph.D. degree from Jiangsu University, Zhenjiang, Jiangsu, China, in 2019. She is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Multimedia Computing Group, 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.

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 has hold 40 granted/pending patents and published over 300 refereed research articles in these areas. He is also an IAPR Fellow and an ACM Distinguished Scientist. He has served as an associate editor, a guest editor, a general chair, a program chair, an area/track chair, a special session organizer, a session chair, and a TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops, including IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing*, *Communications and Applications*, and ACM Multimedia Conference.