Received December 19, 2018, accepted December 24, 2018, date of publication January 10, 2019, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891668

cGAN Based Facial Expression Recognition for Human-Robot Interaction

JIA DENG[1], GAOYANG PANG[1], ZHIYU ZHANG[1], ZHIBO PANG[2], (Senior Member, IEEE), HUAYONG YANG[1], (Member, IEEE), AND GENG YANG[1], (Member, IEEE)

[1]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China
[2]ABB Corporate Research Sweden, 72178 Västerås, Sweden

Corresponding author: Geng Yang (yanggeng@zju.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1301203, in part by the Fundamental Research Funds for the Central Universities, in part by the National Natural Science Foundation of China under Grant U1509204, in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 51521064, in part by the China's Thousand Talents Plan Young Professionals Program, and in part by the Zhejiang University Robotics Institute under Grant K18-508116-009.

ABSTRACT As an emerging research topic for proximity service (ProSe), automatic emotion recognition enables the machines to understand the emotional changes of human beings which can not only facilitate natural, effective, seamless, and advanced human–robot interaction or human–computer interface but also promote emotional health. Facial expression recognition (FER) is a vital task for emotion recognition. However, significant gap between human and machine exists in FER task. In this paper, we present a conditional generative adversarial network-based approach to alleviate the intra-class variations by individually controlling the facial expressions and learning the generative and discriminative representations simultaneously. The proposed framework consists of a generator G and three discriminators (*Di*, *Da*, and *Dexp*). The generator G transforms any query face image into another prototypic facial expression image with other factors preserved. Armed with action units condition, the generator G pays more attention to information relevant to facial expression. Three loss functions (L_I, *La*, and *Lexp*) corresponding to the three discriminators (*Di*, *Da*, and *Dexp*) were designed to learn generative and discriminative representations. Moreover, after rendering the generated expression back to its original facial expression, cycle consistency loss is also applied to guarantee the identity and produce more constrained visual representations. Optimized by combining both synthesis and classification loss functions, the learnt representation is explicitly disentangled from other variations such as identity, head pose, and illumination. Qualitative and quantitative experimental results demonstrate the proposed FER system is effective for expression recognition.

INDEX TERMS Facial expression recognition, emotion recognition, conditional generative adversarial network, human-robot interaction.

I. INTRODUCTION

As an emerging research topic for Proximity Service (ProSe) [1]–[3], safe, natural, and advanced human-robot interaction (HRI) system is supposed to provide not only friendly physical contact between robots and human beings, but also emotional interaction. Among human emotional communication channels, facial expression is arguably the most important visual cue for reflecting the underlying human intentions, physiological changes, affective and cognitive mental states [4]. Therefore, automatic facial expression recognition (FER) pays a vital role in emotional communication based HRI. For instance, as illustrated in Fig. 1, the artificial intelligence agent with soft skins [5] can work alongside people as cooperative teammates to improve productivity. After face detection and deep feature learning, the intelligent robot is able to recognize nuanced meanings conveyed by facial expressions. Moreover, the agent can even improve people's quality of life by taking their emotional health into account in the system and service design. Apart from HRI/human-computer interface (HCI) [10]–[14] and assistive robotics [6]–[9], automatic FER is also important in other applications including movie or advertisement recommendations, driver fatigue surveillance, student engagement estimation [15], and the improvement of expression production in autism disorder patients [16].

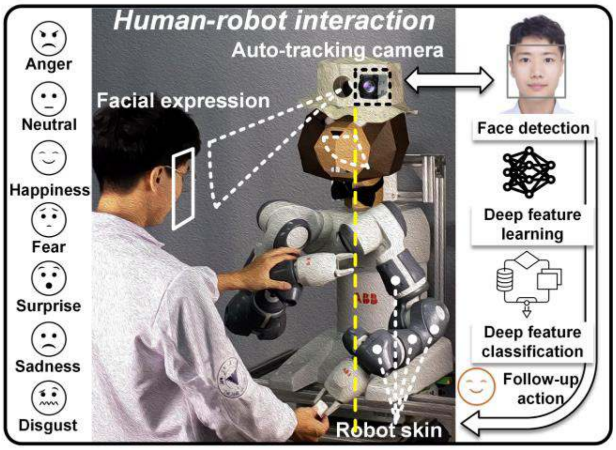

FIGURE 1. Emotional recognition and friendly physical contact based human-robot interaction (HRI). Icons are created by Olena Panasovska and Knut M. Synstad from the Noun Project.

In recent years, great progress has been achieved in automated facial expression analysis on lab controlled datasets which were collected under uniform background and illumination, such as CK+ [17], MMI [18], and Oulu-CASIA [19]. Nevertheless, accurate FER under wild conditions still remains an unsolved issue. The task of FER is challenging because the strong intra-class variability exists due to the various personal identities, such as age, gender, and ethnic backgrounds. Besides the identity bias, other adverse factors in the wild condition may include poor illumination, low resolution, blur, as well as head deflection, etc. Additionally, taken "in-the-wild" datasets are unbalanced. For instance, the training sample number of happy expression images is much bigger than the training sample number of angry expression images in taken 'in-the-wild' datasets [60], [61].

Therefore, in this work, we propose a conditional generative adversarial network (cGAN) based network to disentangle the facial expression factor and learn the generative and discriminative representations simultaneously. Our proposed model consists of a generator G and three discriminators (Di, Da, and $Dexp$). The task of disentangling the facial expression factor is realized in two stages: learning by the conditional generator G and learning by three discriminators. The generator G individually controls the facial expressions by transforming any query face image into another synthetic facial expression image where others factors are preserved such as identity and background, etc. The generator G was also designed for data enrichment to alleviate the imbalance of the data set. Three loss functions (L_I, La, and $Lexp$) were developed corresponding to the three discriminators (Di, Da, and $Dexp$). In addition, after rendering the generated expression back to its original facial expression, cycle consistency loss is also applied to guarantee the identity and produce more constrained visual representations. Optimized by combining both synthesis and classification loss functions, not only the synthesized facial expression images were preserved with identity and background, but also more discriminative features for the expression recognition were obtained.

The rest of this paper is organized as follows. Section II presents three main steps required in a FER system and describes the related background. Section III provides facial image preprocessing, details of proposed neural network architecture and optimization strategy. Adopted facial expression databases, experimental results and analysis of the proposed methodology are introduced in Section IV. Finally, the conclusions are presented in Section V.

II. RELATED WORK

A. FACIAL EXPRESSION RECOGNITION

1) FER APPROACHES BASED ON SHALLOW LEARNING

The majority of existing FER systems focus on six basic emotions types, namely: happy, surprised, fearful, sad, angry, and disgust, which were defined by Ekman [20], [21]. Automatic FER consists of three main stages: pre-processing, facial feature extraction, and expression classification. According to the adopted feature representation, traditional handcrafted FER approaches or so called shallow learning FER approaches can be approximately categorized into four main groups: geometric features based methods, appearance features based methods, action unit (AU) based methods, and motion features based methods.

Appearance-based methods capture global and detailed information by leveraging image filter or filter bank. Pixel intensity [22], Gabor texture [23], local binary patterns (LBP) [24], and histogram of oriented gradients (HOG) [25] are popular descriptors for the appearance-based feature extraction methods. Geometric features based methods work with shapes, positions of the facial components and components' geometric relationships. Motion features are commonly used in video analysis which mostly focus on the temporal correlations of contiguous frames in a sequence, such as motion history images (MHI) [26], volume LBP [27], and optical flow [28]. AU based method was inspired by the physiological and psychological theory. As different facial expressions are the results of the different facial muscles motions, Facial Action Coding System (FACS) [29]–[31] developed by Ekman and Friesen defines facial AUs, the basic elements in formularizing facial expressions, to describe facial muscle activations. Thus facial expressions can be decomposed into multiple AUs. Fig. 2 shows eight basic AUs. Fusion of different handcrafted features was also investigated in previous work. In [32], the texture features and landmark features extracted from facial images were combined which are complementary with each other. Allaert *et al.* [33] considered a hybrid of motion features with geometric features. Additionally, the shape of facial regions of interest were exploited to form the apex frame [33].

2) FER APPROACHES BASED ON DEEP LEARNING

Over the last few years, deep convolutional neural networks (CNNs) [34]–[36] have produced unprecedented

FIGURE 2. Action unit images for AU 1, 4, 6, 9, 12, 15, 26, and 28.

performance on a variety of tasks, such as object recognition [37], scene classification [38], and face recognition [39]–[41]. Meanwhile, conventional handcrafted features or shallow learning based FER approaches have been reported a limited recognition performance, because they lack the ability to cope with the great diversity of factors which are irrelevant to facial expressions. These factors irrelevant to facial expressions include backgrounds, hair, and head deflection. Consequently, utilization of deep learning techniques in FER has attracted considerable attention among researchers. Yang *et al.* [42] leveraged a partial VGG16 network and a shallow CNN to extract two feature vectors from facial grayscale images and LBP facial images, respectably. Then the two feature vectors were fused to fully use complementary facial information. Tang *et al.* [43] extracted features by twelve convolutional and pooling layers which were more efficient and provided a great improvement, compared with the 78 dimensions geometric features. Zheng *et al.* [44] presented a VGG16 + 1D-CNN model for FER. In their framework, representations of each frame of a video were extracted with VGG16 network followed by four 1D-CNN networks. Then the features were concatenated and were fed to two fully connected (FC) layers to predict facial expressions.

B. GENERATIVE ADVERSARIAL NETWORKS (GANS)

Generative adversarial networks (GANs) [45] have been vigorously studied in recent years, since the model has achieved remarkable results in various computer vision tasks such as image generation, image translation, etc. The goal of GANs is to model distribution as similar as possible to the true data distribution. To achieve this goal, the generator G and the discriminator D of GANs compete in a two player minimax game. Specifically, the discriminator learns to distinguish real samples from fake samples while the generator learns to generate fake samples to fool the discriminator until reach the Nash equilibrium [46] between the two modules. The cGAN [47] is an extension of the GAN where the model receives additional variables (features or label, etc.) as input, which could deterministically control the output of the generator. cGAN has been successfully applied to synthesize images from labels, reconstructing objects from edge maps, and photo editing [48], etc. Lai and Lai [49] proposed a GAN-based network model to achieve canonical-view facial expression recognition. The generator in their model frontalized input non-frontal face images into frontal face images while preserving the identity and expression characteristics. Yang *et al.* [50] presented a cGAN based approach which generated facial expressions in order to alleviate the issue of subject variations. The input of their model was restricted to image pairs where each image pair included two different expressions of the same person. However, paired images are usually not available in the wild condition. For instance, there are rarely paired images in AffectNet [60] and RAF-DB [61] datasets which are taken in the wild. In contrast, the proposed model in this paper is capable of dealing with unpaired data. As AUs are the basic elements of facial expressions, we present an EAU-Net network to transform any query face image into another prototypic facial expression image by editing AUs. More specifically, with desired AUs condition, the cGAN based EAU-Net edits original AUs of a given face image and reconstructs a synthetic face image with desired AUs. In the meantime, generative and discriminative representations are learnt for recognition. To this end, two different expressions of the same person are dispensable.

III. PROPOSED METHOD

A. FACIAL IMAGE PRE-PROCESSING

Adverse variations exist in wild condition, such as complex backgrounds and poor illumination, etc. Therefore, pre-processing to align and normalize the facial images is necessary, before deep feature learning. The three steps for facial image pre-processing are described below. Step1: Crop the face region to remove the uncorrelated information. To crop the facial image, firstly, multi-task cascade convolutional neural networks (MTCNN) [51] is employed to detect face and to provide the bounding box of facial region, as MTCNN is found to be robust and effective for alignment [52]. Then, according to the bounding box, the face region is cropped from the original facial image. Step2: Resize the cropped image to a fixed size which makes sure that the same scale is shared among all images. In order to capture more subtle facial expression information, the fixed size is set to 256×256 pixels. Step3: Normalize the resized images from [0, 255] to [0, 1] and augment the data using techniques such as random flip. The purpose of normalizing the image inputs is not only to remove the high frequency noise but at the same time to ensure that the pixels of the image have a similar distribution. Additionally, after random flipping (flipping an image horizontally or vertically), the data sample size is expanded, which is very helpful to improve the accuracy and the generalization capability of the model.

B. NETWORK ARCHITECTURE

A cGAN based network for data enrichment whilst performing FER is proposed in this paper. The overall architecture

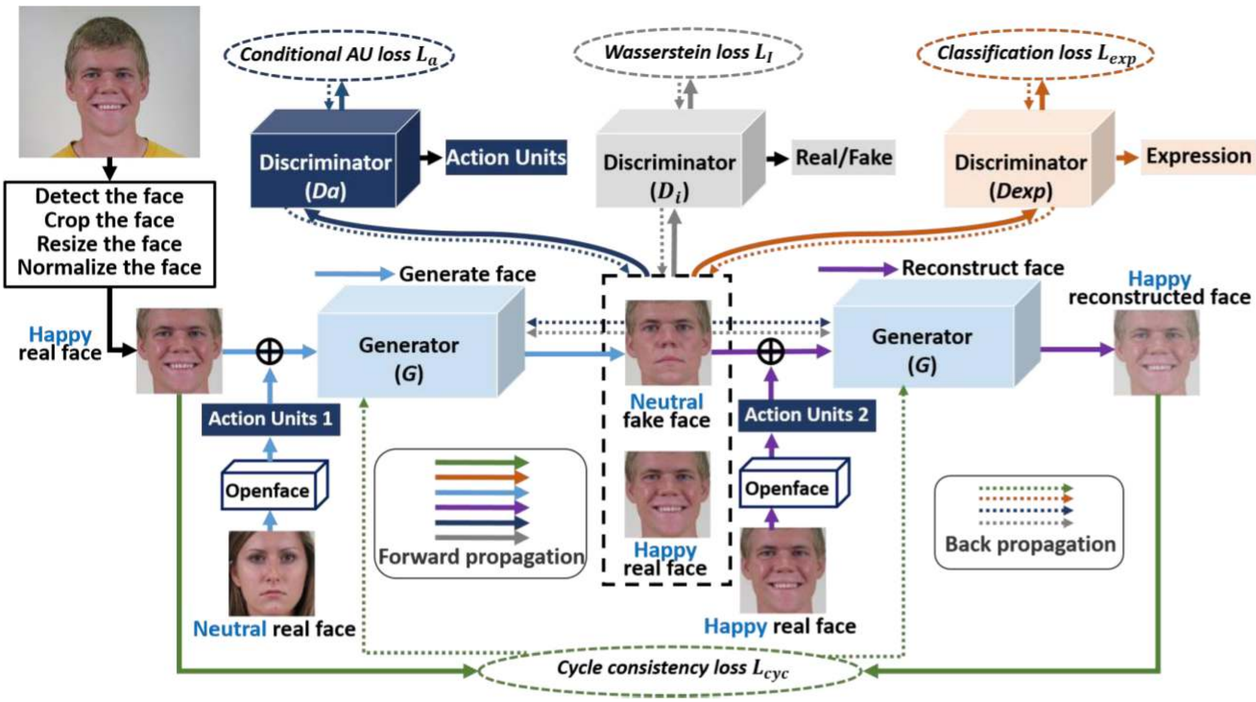

FIGURE 3. The overall architecture of the proposed model, which consists of a generator *G* and three discriminators (*Di*, *Da*, and *Dexp*). The generator *G* is applied twice: *G* transforms any given query face image to another facial expression image and then rends it back. Openface [53] is applied to extract the AUs.

is depicted in Fig. 3, which composes of a generator *G* and three discriminators (*Di*, *Da*, and *Dexp*). The processing flow is described as follows:

Here we take the transformation of a happy real face to a neutral fake face as an example, as shown in the lower part of Fig. 3. Firstly, a cropped, resized, normalized and augmented facial expression image (happy real face) concatenated with AUs is fed to the generator *G* to generate a synthetic face (neutral fake face). The AUs are a set of 17-dimensional vectors, and AUs applied here are extracted from a neutral real face using Openface [53], as shown in Fig. 3. Actually, the proposed generator *G* is capable of transforming any given facial expression image into another prototypic facial expression (happy, surprised, fearful, sad, angry, disgust, or neutral) image. For instance, with the AUs extracted from a sad real face, the generator *G* is able to transform any given facial expression image to a sad fake face.

Secondly, both the real face and fake face are sent to three discriminators. The functions of the three constructed discriminators vary, and the specific functions of each discriminator are described as below: 1) Discriminator *Di*: In order to generate photorealistic image, the discriminator *Di* is indispensable. The discriminator *Di* evaluates the quality of the generated image to distinguish the real face from the fake face. 2) Discriminator *Da:* The discriminator *Da* learns to estimate the AUs values to make sure that the AUs from the neutral fake face is similar to the AUs from the neutral real face. 3) Discriminator *Dexp*: The discriminator*Dexp* performs the FER task which learns to predict facial expression labels (including happy, surprised, fearful, sad, angry, disgust, and neutral).

Thirdly, the generator *G* renders the generated face (neutral fake face) back to the original facial expression (happy) to generate a reconstructed face (happy reconstructed face). This reconstruction guarantees the generator *G* mapping one facial expression to another facial expression with identity (including age, gender, and ethnic backgrounds, etc.) and other factors (including background and illumination, etc.) preserved by minimizing cycle consistency loss L_{cyc}. The cycle consistency loss L_{cyc} is defined as the difference between the original image (happy real face in this case) and the reconstructed image (happy reconstructed face in this case).

Finally, the parameters of generator *G* and three discriminators (*Di*, *Da*, and *Dexp*) are learnt by optimizing four loss functions (L_I, *La*, *Lexp*, and *Lcyc*). The details of these four loss functions are described in part C of section III.

The detailed structures of each component in the proposed model are described below: 1) Generator *G*: Fig. 4 reports the architecture of the proposed generator *G* where convolutional encoder-decoder layers are embedded. More specifically, the generator *G* comprises an encoder with output channels {64, 128, 256} and a decoder with output channels {128, 64, 3}. Batch normalization (BN) [54] is applied between each convolutional layer ("Conv3" or "Conv4")/ deconvolutional layer ("Deconv3" or "Deconv4") and non-linear activation function (ReLU). The function of BN here is to reduce internal covariate shift to regularize the model and to improve the convergence speed. As shown

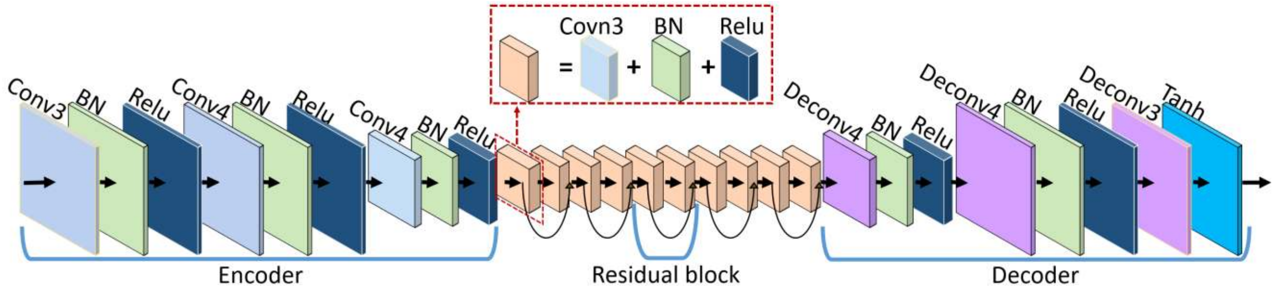

FIGURE 4. The architecture of the proposed generator G where convolutional encoder-decoder layers are embedded.

TABLE 1. A Detailed Description of The Architecture of the Proposed Discriminator Di. The Output Shape is Described as (#channels, height, width).

Layer Type		Kernel	Output	Stride	Pad
1	Conv1	4	64*128*128	2	1
2	ReLU1	-	-	1	0
3	Conv2	4	128*64*64	2	1
4	ReLU2	-	-	1	0
5	Conv3	4	256*32*32	2	1
6	ReLU3	-	-	1	0
7	Conv4	4	512*16*16	2	1
8	ReLU4	-	-	1	0
9	Conv5	4	1024*8*8	2	1
10	ReLU5	-	-	1	0
11	Conv6	4	2048*4*4	2	1
12	ReLU6	-	-	1	0
13	Conv7	4	17*1*1	1	0

TABLE 2. A Detailed Description of The Architecture of the Proposed Discriminator Da. The Output Shape is Described as (#channels, height, width).

Layer Type		Kernel	Output	Stride	Pad
1	Conv1	4	64*128*128	2	1
2	ReLU1	-	-	1	0
3	Conv2	4	128*64*64	2	1
4	ReLU2	-	-	1	0
5	Conv3	4	256*32*32	2	1
6	ReLU3	-	-	1	0
7	Conv4	4	512*16*16	2	1
8	ReLU4	-	-	1	0
9	Conv5	4	1024*8*8	2	1
10	ReLU5	-	-	1	0
11	Conv6	4	2048*4*4	2	1
12	ReLU6	-	-	1	0
13	Conv7	3	1*4*4	1	1

in Fig. 4, "Conv3" and "Conv4" employ 3×3 filters and 4×4 filters, respectively. Similarly, "Deconv3" and "Deconv4" employ 3×3 filters and 4×4 filters, respectively. Residual blocks are also used here, where the stride size is set as 1 and the filter size is set as 3×3. 2) Discriminator *Di* and Discriminator *Da*: The detailed descriptions of the architectures of the proposed discriminator *Di* and discriminator *Da* are provided in Table 1 and Table 2, respectively. Both the discriminator *Di* and the discriminator *Da* have a 13-layers structure. For instance, layer 1 of discriminator *Di* is a convolutional layer with filter size of 4×4 where the stride and the pad are set to 2 and 1, respectively. And the outputs of layer 1 are $64 \times 128 \times 128$ feature maps. The architectures of the two discriminators (*Di* and *Da*) are similar. The only difference between the two discriminators is the last convolutional layer. In the discriminator *Di*, the last convolutional layer employs 4×4 filters and provides $17 \times 1 \times 1$ feature maps as output where the stride and the pad are set to 1 and 0, respectively. In the discriminator *Da*, $1 \times 2 \times 2$ feature map is outputted from the last convolutional layer with filter size of 3×3 where the stride and the pad are set to 1 and 1, respectively. 3) Discriminator*Dexp*: VGGNet-19 network [70] is applied in the discriminator *Dexp* which is trained using original images and synthetic images generated from generator G. The detailed description of VGGNet-19 model is illustrated in Table 3. It can be observed that the VGGNet-19 model follows a conventional CNN structure, comprising 16 convolutional layers and 3 fully connected layers. The VGGNet-19 model also includes 5 pooling layers which are used to reduce the number of parameters to speed up the computation.

C. OPTIMIZATION STRATEGY

1) ADVERSARIAL LOSS (L_{adv})

The generative network (generator) G and the discriminative network (discriminator) D compete in a two player mini-max game. In the game, the generator G generates synthetic images to fool the discriminator D while the discriminator D in turn tries to accurately distinguish the real images from the generated images. Given the training data, GAN is trained by optimizing the adversarial objective $\min_G \max_D L_{adv}$. The discriminator D tries to maximize the adversarial loss while the generator G tries to minimize it. The adversarial loss is

TABLE 3. A Detailed Description of The VGGNet-19 Architecture. The Output Shape is Described as (#channels, height, width).

Layer Type		Kernel	Output	Stride	Pad
1	Conv1-1	3	64*224*224	1	1
2	ReLU1-1	-	-	1	0
3	Conv1-2	3	64*224*224	1	1
4	ReLU1-2	-	-	1	0
5	Pool1	2	64*112*112	2	0
6	Conv2-1	3	128*112*112	1	1
7	ReLU2-1	-	-	1	0
8	Conv2-2	3	128*112*112	1	1
9	ReLU2-2	-	-	1	0
10	Pool2	2	128*56*56	2	0
11	Conv3-1	3	256*56*56	1	1
12	ReLU3-1	-	-	1	0
13	Conv3-2	3	256*56*56	1	1
14	ReLU3-2	-	-	1	0
15	Conv3-3	3	256*56*56	1	1
16	ReLU3-3	-	-	1	0
17	Conv3-4	3	256*56*56	1	1
18	ReLU3-4	-	-	1	0
19	Pool3	2	256*28*28	2	0
20	Conv4-1	3	512*28*28	1	1
21	ReLU4-1	-	-	1	0
22	Conv4-2	3	512*28*28	1	1
23	ReLU4-2	-	-	1	0
24	Conv4-3	3	512*28*28	1	1
25	ReLU4-3	-	-	1	0
26	Conv4-4	3	512*28*28	1	1
27	ReLU4-4	-	-	1	0
28	Pool4	2	512*14*14	2	0
29	Conv5-1	3	512*14*14	1	1
30	ReLU5-1	-	-	1	0
31	Conv5-2	3	512*14*14	1	1
32	ReLU5-2	-	-	1	0
33	Conv5-3	3	512*14*14	1	1
34	ReLU5-3	-	-	1	0
35	Conv5-4	3	512*14*14	1	1
36	ReLU5-4	-	-	1	0
37	Pool5	2	512*7*7	2	0
38	FC6	-	4096	-	-
39	ReLU6	-	-	1	0
40	FC7	-	4096	-	-
41	ReLU7	-	-	1	0
42	FC8	-	1000	-	-

defined in equation (1) below:

$$\boldsymbol{L_{adv}} = (\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^{x}}\left[\boldsymbol{logD}\left(\boldsymbol{x};\boldsymbol{\theta_D}\right)\right] + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^{x}}\left[\boldsymbol{log}(1-\boldsymbol{D}\left(\boldsymbol{G}(\boldsymbol{x};\boldsymbol{\theta_G});\boldsymbol{\theta_D}\right)\right] \quad (1)$$

where x is the input image from the training data, $\mathbb{P}^x$ denotes the distribution of the training data, $\mathbb{E}[\cdot]$ represents the expected value operator, θ_D and θ_G are the parameters of discriminator D and generator G, respectively.

2) WASSERSTEIN LOSS (L_I)

However, training GAN with L_{adv} is unstable [55]. Thus, one of the most stable variation of GAN called Wasserstein GAN (WGAN) [56] is employed in this work. WGAN allows a stable training of GAN by minimizing an approximation of the Wasserstein distance [57] which is an efficient metric to measure the dissimilarity between two multidimensional data sets. The Wasserstein loss in this case is calculated as follows:

$$\begin{aligned} L_I = (&\mathbb{E}_{x_{a_o}\sim\mathbb{P}^x}\left[D_i\left(x_{a_o};\theta_{D_i}\right)\right] \\ &- \mathbb{E}_{x_{a_o}\sim\mathbb{P}^x}\left[D_i\left(G(x_{a_o}|a_T;\theta_G);\theta_{D_i}\right)\right] \\ &- \lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}^x_\alpha}\left[\left(\|\nabla_{\hat{x}}D_i\left(\hat{x};\theta_{D_i}\right)\|_2 - 1\right)^2\right] \end{aligned} \quad (2)$$

where x_{a_o} is the input image with AUs a_0, $\mathbb{E}[\cdot]$ represents the expected value operator, a_T denotes the target AUs, λ is the penalty coefficient, ∇ represents the vector differential operator, θ_{D_i} and θ_G are the parameters of discriminator Di and generator G, respectively, $\mathbb{P}^x$ is the distribution of the training data, $\mathbb{P}^x_\alpha$ is the joint distribution of the original images and the synthetic images $G(x_{a_o}|a_T;\theta_G)$ produced by generator G, and $\hat{x}$ is defined as $\hat{x} = \alpha x + (1-\alpha)G(x_{a_o}|a_T;\theta_G)$, with $\alpha \sim U(0, 1)$ (i.e., uniform distribution).

3) CONDITIONAL AUS LOSS (L_a)

With the AUs condition, the generator G maps any query face image to another prototypic facial expression image according to the regions relevant to facial expression. Thus, the discriminator Da is employed to estimate the AUs values which forces the generator G to make its best efforts to generate more nuanced facial expression images. As AUs can be extracted from the generated images as well as the original images, the conditional AUs loss is defined in equation (3) below:

$$\begin{aligned} L_a = \mathbb{E}_{x_{a_o}\sim\mathbb{P}^x}&\left[\|D_a\left(G\left(x_{a_o}\,|\,a_T;\theta_G\right);\theta_{D_a}\right) - a_T\|_2^2\right] \\ &+ \mathbb{E}_{x_{a_o}\sim\mathbb{P}^x}\left[\|D_a\left(x_{a_o};\theta_{D_a}\right) - a_o\|_2^2\right] \end{aligned} \quad (3)$$

where x_{a_o} is the input image with AUs a_0, $\mathbb{E}[\cdot]$ represents the expected value operator, a_T denotes the target AUs, $\mathbb{P}^x$ is the distribution of the training images, $G\left(x_{a_o}\,|\,a_T;\theta_G\right)$ represents the generated image, θ_{D_a} and θ_G are the parameters of discriminator Da and generator G, respectively.

4) CYCLE CONSISTENCY LOSS (L_{cyc})

Although the GAN based FER approaches were investigated in previous studies [50], [58], the unconstrained nature of the mapping process (from one facial expression to another facial expression) may produce distribution that is far away from the real distribution in the training set, resulting in an ineffective multi-class classifier training. Therefore, a multi-modal cycle consistency loss is adopted. The cycle consistency loss is used to estimate the reconstruction error between the original facial expression image and the reconstructed facial expression image. This regularization is inspired by cycle consistency loss [59]. After rendering the generated expression back to its original facial expression, the cycle consistency loss is applied in training GAN to produce more constrained visual representations to maximally maintain identity information and other factors (including background and illumination, etc.). The above-mentioned cycle consistency loss is computed as:

$$L_{cyc} = (\mathbb{E}_{x_{a_o} \sim \mathbb{P}^x} \left[\| \, G \left(G \left(x_{a_o} \,|\, a_T; \theta_G \right) \,|\, a_o; \theta_G \right) - x_{a_o} \, \|_1 \right] \tag{4}$$

where x_{a_o} is the input image with AUs a_0, $\mathbb{E}[\cdot]$ represents the expected value operator, a_T denotes the target AUs, θ_G are the parameters of generator G, $\mathbb{P}^x$ is the distribution of the training images, $G\left(x_{a_o} \,|\, a_T; \theta_G\right)$ represents the generated image, $G\left(G\left(x_{a_o} \,|\, a_T; \theta_G\right) \,|\, a_o; \theta_G\right)$ denotes the reconstructed image.

5) CLASSIFICATION LOSS (L_{exp})

While the WGAN loss and the conditional AUs loss are applied, they do not guarantee that the generated facial images are discriminative for facial expression classification. Consequently, the classification loss is formulated. The classification loss encourages the generator G to construct images that can be correctly categorized into different facial expression labels by discriminator *Dexp*. The classification loss for the expression classification is defined as:

$$L_{exp} = -\mathbb{E}_{x \sim \mathbb{P}^x} \left[\log D_{exp}(y|x; \theta_{D_{exp}}) \right] \tag{5}$$

where x is the input image, $\mathbb{P}^x$ denotes the distribution of the training data, $\mathbb{E}[\cdot]$ represents the expected value operator, $\theta_{D_{exp}}$ are the parameters of discriminator *Dexp*, the term $D_{exp}(y|x; \theta_{D_{exp}})$ represents a probability distribution over expression labels computed by *Dexp*.

6) FULL LOSS (L)

Finally, we use the full loss function by combining the four loss functions (L_I, L_a, L_{cyc}, and L_{exp}):

$$L = \lambda_1 L_I + \lambda_2 L_a + \lambda_3 L_{cyc} + \lambda_4 L_{exp} \tag{6}$$

where λ_1, λ_2, λ_3, and λ_4 are hyper-parameters for adjusting the weights of individual loss functions. Actually, during the course of the experiment, we find that dividing the training process into two phases is beneficial in terms of improving training stability and speeding up the convergence. More specifically, in the first phase of the two-phase scheme, $\lambda 1$, $\lambda 2$, $\lambda 3$, and $\lambda 4$ are set to 1, 4000, 10, and 0, respectively. In the second phase of the two-phase scheme, $\lambda 1$, $\lambda 2$, $\lambda 3$, and $\lambda 4$ are set to 0, 0, 0, and 1, respectively. These weight hyper-parameters are chosen through numerous experiments.

IV. EXPERIMENTS AND DISCUSSION

A. IMPLEMENTATION DETAILS

AffectNet [60] and Real-world Affective Faces Database (RAF-DB) [61] taken "in-the-wild" datasets are used, since these two datasets are more approximate to the real world scenarios than posed datasets collected in a constrained laboratory. Though several other databases such as CK+ [17], MMI [18], Oulu-CASIA [19], and JAFFE [62] for FER are available, most of them are sampled in well controlled environment. The details of AffectNet and RAF-DB datasets are provided in Table 4.

TABLE 4. The Details of Experiment Set for FER including the Expression Categories, Training, and Validation Samples.

Dataset	AffectNet		RAF-DB	
Tr.\Val. Sample	Training Samples	Validation Samples	Training Samples	Validation Samples
Happy	87695	20237	4772	1185
Sad	16873	3894	1982	478
Surprise	9484	2188	1290	329
Fearful	4471	1031	281	74
Disgust	2797	645	717	14545
Angry	16498	3808	705	162
Neutral	48993	11306	2524	680
Total	186811	43109	12271	3068

To date, AffectNet is the largest database with annotated facial emotions [60]. It contains about 400,000 images and each image is labeled with one of the discrete facial expressions (including neutral, anger, disgust, fear, happy, sad, surprise, and contempt). Nevertheless, limitations exist in AffectNet database. For instance, each image is annotated by only one labeler. Following [71], [72], around 280,000 images with seven prototypic facial expressions (anger, disgust, fear, happy, sad, surprise, and neutral) are selected as training samples and 3,500 images as validation samples in this work.

RAF-DB is a large-scale facial expression database with around 30,000 great-diverse facial images downloaded through various search engines [61]. The images in this dataset vary in personal identities (including age, gender, and ethnic backgrounds, etc.), head pose, and lighting conditions, etc. And each image from RAF-DB dataset contains more annotation information which is the effort result of about 40 independent labelers, compared with the images from AffectNet dataset. In RAF-DB dataset, 15331 images are labeled with seven basic expression categories (anger, disgust, fear, happy, sad, surprise, and neutral) where 12271 are used for training and 3068 for validation.

The implementation is carried on the workstation accelerated by GeForce GTX 1080Ti 11G. And the EAU-Net model is developed in the deep learning framework Pytorch [63]. Training a single two-phase proposed model EAU-Net takes 5.2 hours for 20k iterations with the batch size of 16 on RAF-DB dataset while the one-phase EAU-Net takes 8 hours for training. And it takes 50 hours to train a single two-phase EAU-Net model for 30 epochs with the batch size of 48 on AffectNet dataset.

TABLE 5. Expression Recognition Accuracies of Different Methods on the RAF-DB Database.

Method	Training Samples	Validation Samples	Average Accuracy
Transfer ResNet [64]	12271	3068	80%
CapsNet [65]	12271	3068	77.48%
AUG-CLOSS [66]	12271	3068	75.73%
Double Cd-LBP [67]	12271	3068	78.6%
DLP-CNN [68]	12271	3068	74.2%
MRE-CNN [69]	12271	3068	76.73
Proposed (EAU-Net)	12271	3068	81.83%

B. QUANTITATIVE EVALUATIONS OF THE PROPOSED APPROACH

For the FER task on RAF-DB database, the proposed method performance is compared with performances of recently published methods in literatures [64]–[69], as shown in Table 5. It can be seen that the method proposed in this paper achieves an average recognition accuracy up to 81.83%, which outperforms the listed state-of-the-arts methods including handcrafted feature based method [67], CNN-based methods [64], [68], [69], capsule-based method [65], and data augmentation based method [66]. Compared with the listed state-of-the-arts methods, our cGAN based approach is able to disentangle facial expression factor by individually controlling the facial expressions and optimizing both synthesis and classification loss functions (L_I, La, $Lcyc$, and $Lexp$) and thus achieves high accuracy in FER task on RAF-DB database.

TABLE 6. Expression Recognition Accuracies of Different Methods on the AffectNet Database.

Method	Training Samples	Validation Samples	Average Accuracy
VGG [70]	283900	3500	51.11%
IPA2LT [71]	283900	3500	56.51%
PG-CNN [72]	283900	3500	55.33%
Proposed (EAU-Net)	283900	3500	58.91%
VEGAC [73]	186811	43109	72.2%
Proposed (EAU-Net)	186811	43109	74.80%

Table 6 shows the comparison between our work and other state-of-the-arts methods for the FER task on AffectNet database. Among these methods, [70], [71] are CNN based approaches and [72], [73] are CNN with attention based approaches. The model proposed in this paper achieves 74.80% accuracy for the FER task evaluated on AffectNet database which outperforms the listed state-of-the-arts methods. In [73], visual salient regions joined with original face image were fed to CNN to perform FER while the visual salient regions were just found to be more related to eyes, mouth, and nose, these rough regions. However, our cGAN based model is capable of not only learning the regions related to expression, but also maximally capturing nuanced characteristics relevant to expression and then transforming the original expression to another expression with identity and other factors preserved which is shown in Fig. 7 and Fig. 8.

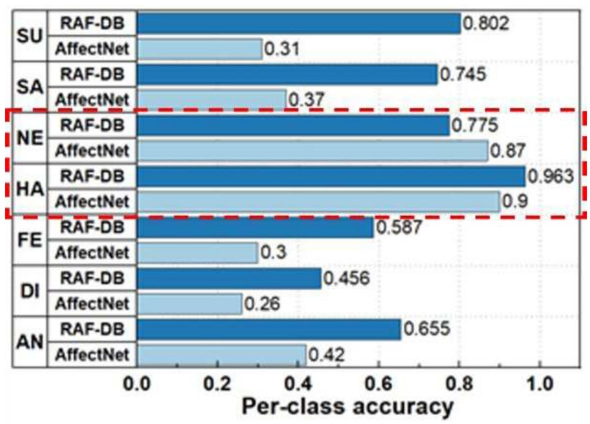

FIGURE 5. Per-class accuracy of RAF-DB dataset and AffectNet dataset where the seven facial expression classes include surprise (SU), sad (SA), neutral (NE), happy (HA), fear (FE), disgust (DI), and anger (AN).

Per-class accuracy of RAF-DB dataset and AffectNet dataset is illustrated in Fig. 5 where the seven facial expression classes include surprise (SU), sad (SA), neutral (NE), happy (HA), fear (FE), disgust (DI), and anger (AN). It can be seen that the top two with the highest recognition rates in both datasets are NE and HA. However, the classification accuracies of SU, FE, DI and SA expressions evaluated on AffectNet dataset are relatively low compared with other expressions. Reason is the training sample numbers of SU, FE and DI expressions are much fewer than others which are shown in Table 4. Additionally, in comparison with the facial expression images in RAF-DB database, the facial expression images in AffectNet dataset are more difficult to distinguish, even by humans. Examples can be found in Fig. 6; four expression samples (surprise, fear, disgust, and sad) from AffectNet dataset have only tiny difference and these expressions can be easily confused with each other resulting in poor recognition performance. Overall, the high classification accuracies evaluated on both RAF-DB and AffectNet datasets indicate that the proposed network is effective for the facial expression classification task in wild conditions.

FIGURE 6. Ambiguous samples of the four facial expression classes (sad, disgust, fearful, and surprise) from AffectNet dataset.

C. QUALITATIVE EVALUATIONS OF THE PROPOSED APPROACH

Some qualitative results are visualized in Fig. 7 and Fig. 8. In Fig. 7, the left column are the real images and the right

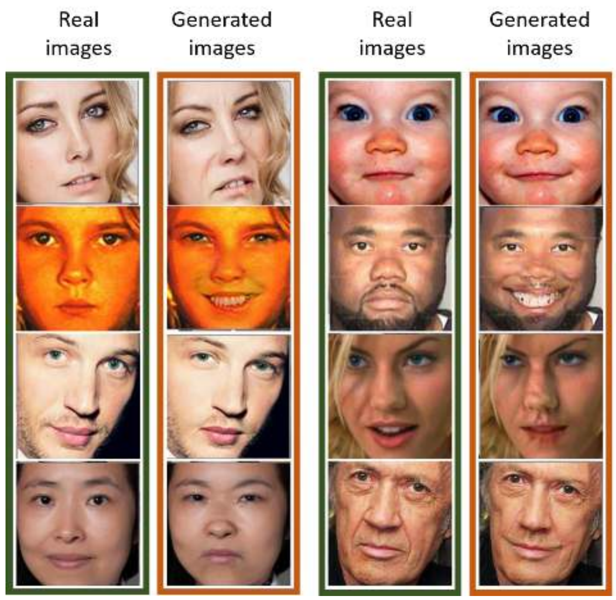

FIGURE 7. Synthesis results: real facial image (left column), generated facial image (right column).

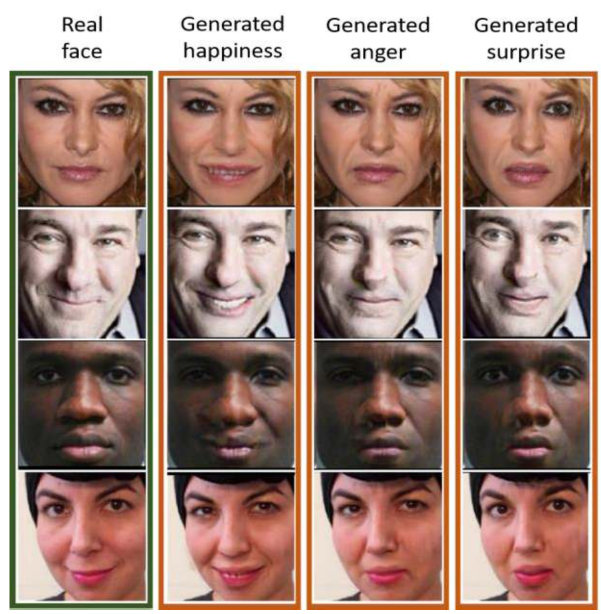

FIGURE 8. Synthesis results: real facial image (left column), three generated facial images (right three columns).

column are the generated images. It can be observed that the generated images are disentangled from other variations such as backgrounds, head pose, and illumination, etc. It vividly shows that the feasibility of individually controlling the facial expressions and simultaneously learning of the generative and discriminative representations. In addition, the proposed model achieves high-quality image synthesis results, even in the cases with various light intensity and head deflection. As shown in the right columns of Fig. 7, the details of facial characteristics like hair, skin color, wrinkles, background and illumination are generated nicely. In Fig. 8, the left column are the real images and the right three columns are generated images. Each image is constructed to three synthesis images with expressions of happiness, anger, and surprise, respectively. It demonstrates that our cGAN based model is capable of disentangling the facial expression factor and transforming any given query face image into several images at the same time, each with a different expression.

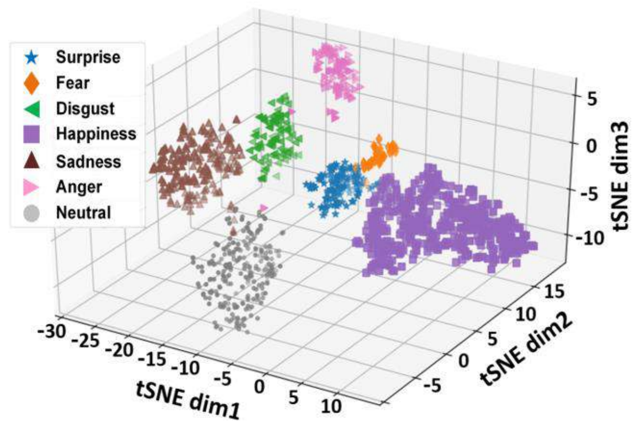

FIGURE 9. The 3-D t-SNE plot of the deep features learnt by the proposed model on RAF-DB dataset.

To visualize the learnt deep features of each expression, T-distributed Stochastic Neighbor Embedding (t-SNE) [74] is employed to nonlinear reduce the learnt high-dimensional features to a three-dimensional space. Fig. 9 presents a 3-D t-SNE plot of the deep features learnt from RAF-DB dataset. The deep features were the output of the last fully connected layer of discriminator *Dexp*. The random sample number is set to 1250 with considering of the computation speed and the number of the validation set of RAF-DB dataset. It can be observed that the dots of fear expression in the 3-D t-SNE plot are relatively few. That is because the data distribution of the publicly available training set of RAF-DB database is unbalance. As it can be seen in Table 4, the available training number of fear expression in the RAF-DB database is only 281 while the training number of happy expression is 4772. Although adverse variations (including various race, age, head pose, and illumination, etc.) exist in RAF-DB dataset, the dots of every expression tend to cluster and there is relatively clear interval among seven expressions which demonstrates the effectiveness of the learnt representations.

V. CONCLUSION

As an emerging research topic for ProSe, automatic FER has attracted a great amount attention in recent years, as FER plays a vital part in emotion recognition and has a variety of applications in HRI and emotion healthcare, etc. However, accurate FER in real world scenarios remains a challenging task due to the complex backgrounds, various light intensity and head deflection, etc. In this paper, we propose a cGAN based approach to disentangle the facial expression factor and learn the generative and discriminative representations simultaneously. The task of disentangling the facial expression factor is implemented in two stages: learning by a conditional generator G and learning by three discriminators (Di, Da, and Dexp). The generator G disentangles the facial expression factor by transforming any face image into a synthetic image with one of the seven basic facial expressions (anger, disgust, fear, happy, sad, surprise, and neutral). Three loss functions (L_I, La, and Lexp) corresponding to three discriminators (Di, Da, and Dexp) are developed to learn the generative and discriminative representations simultaneously. Additionally, the cycle consistency loss is also applied to guarantee that the personal identity, background, head deflection and illumination are persevered. By optimizing the overall loss functions, the learnt representations are disentangled from other variations. The experimental results show that the proposed approach is effective for FER task and the proposed approach outperforms the known competing methods on both AffectNet and RAF-DB datasets. One limitation of this work is that the model is trained individually for different datasets. Since it is subjective to annotate the face expressions, the bias of annotations is inevitable among different datasets. Thus, one model trained on a specific dataset may get poor performance on another dataset with a different distribution for FER task. Making pseudo annotations and learning latent features are worth to be investigated in cross-dataset learning where the training data and the verification data are from different datasets.

REFERENCES

[1] J. Qi, P. Yang, M. Hanneghan, S. Tang, and B. Zhou, "A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2018.2846359.

[2] P. Yang *et al.*, "Lifelogging data validation model for Internet of Things enabled personalized healthcare," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 1, pp. 50–64, Jan. 2018.

[3] J. Qi, P. Yang, A. Waraich, Z. Deng, Y. Zhao, and P. Yang, "Examining sensor-based physical activity recognition and monitoring for healthcare using Internet of Things: A systematic review," *J. Biomed. Inform.*, vol. 87, pp. 138–153, Nov. 2018.

[4] J. F. Cohn, "Foundations of human computing: facial expression and emotion," in *Proc. 8th Int. Conf. Multimodal Interfaces (ICMI)*, Banff, AB, Canada, 2006, pp. 233–238.

[5] G. Pang, J. Deng, F. Wang, J. Zhang, Z. Pang, and G. Yang, "Development of flexible robot skin for safe and natural human–robot collaboration," *Micromachines*, vol. 9, no. 11, pp. 576–591, Nov. 2018.

[6] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *Proc. 9th Int. Conf. Rehabil. Robot. (ICORR)*, Chicago, IL, USA, Jun./Jul. 2005, pp. 465–468.

[7] A. Guler *et al.*, "Human joint angle estimation and gesture recognition for assistive robotic vision," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, 2016, pp. 415–431.

[8] E. S. John, S. J. Rigo, and J. Barbosa, "Assistive robotics: Adaptive multimodal interaction improving people with communication disorders," *IFAC-PapersOnLine*, vol. 49, no. 30, pp. 175–180, Nov. 2016.

[9] A. Bolotnikova *et al.*, "A circuit-breaker use-case operated by a humanoid in aircraft manufacturing," in *Proc. 13th IEEE Conf. Autom. Sci. Eng (CASE)*, Xi'an, China, Aug. 2017, pp. 15–22.

[10] Y. Dai *et al.*, "An associate memory model of facial expressions and its application in facial expression recognition of patients on bed," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Tokyo, Japan, Aug. 2001, pp. 591–594.

[11] G. Anbarjafari and A. Aabloo, "Expression recognition by using facial and vocal expressions," in *Proc. 25th Int. Conf. Comput. Linguistics*, Dublin, Ireland, 2014, pp. 103–105.

[12] I. Lüsi *et al.*, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May/Jun. 2017, pp. 809–813.

[13] K. Nasrollahi *et al.*, "Deep learning based super-resolution for improved action recognition," in *Proc. Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Orleans, France, Nov. 2015, pp. 67–72.

[14] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Madison, WI, USA, Jun. 2003, p. 53.

[15] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan./Mar. 2016.

[16] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *Proc. Workshop Facial Bodily Expressions Control Adaptation Game (ECAG)*, Enschede, The Netherlands, 2008, pp. 978–986.

[17] P. Lucey *et al.*, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Vis. Pattern Recogit. Workshop (CVPRW)*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[18] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the mmi facial expression database," in *Proc. 3rd Int. Workshop EMOTION (Satellite LREC), Corpora Res. Emotion Affect*, Paris, France, 2010, pp. 65–70.

[19] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[20] P. Ekman, *Pictures of Facial Affect*. Palo Alto, CA, USA: Consulting Psychologists Press, 1976.

[21] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, 1993.

[22] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "PCA-based dictionary building for accurate facial expression recognition via sparse representation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1082–1092, Jul. 2014.

[23] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[24] 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, 2009.

[25] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[26] M. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *Proc. IEEE Int. Conf. Syst., Man. Cybern.*, Oct. 2004, pp. 635–640.

[27] G. Zhao and M. Pietikäinen, "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.

[28] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans. Inf. Syst.*, vol. 74, no. 10, pp. 3474–3483, Oct. 1991.

[29] P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System*. Palo Alto, CA, USA: Consulting Psychologists Press, 1977.

[30] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford, U.K.: Oxford Univ. Press, 1997.

[31] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[32] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, Oct. 2015.

[33] B. Allaert, I. M. Bilasco, and C. Djeraba. (May 2018). "Advanced local motion patterns for macro and micro facial expression recognition." [Online]. Available: https://arxiv.org/abs/1805.01951

[34] 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.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[37] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 487–495.
[39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[40] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.
[41] A. Polyak and L. Wolf, "Channel-level acceleration of deep face representations," *IEEE Access*, vol. 3, pp. 2163–2175, 2015.
[42] B. Yang, J. M. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.
[43] Y. Tang, X. M. Zhang, and H. Wang, "Geometric-convolutional feature fusion based on learning propagation for facial expression recognition," *IEEE Access*, vol. 6, pp. 42532–42540, 2018.
[44] Z. Zheng, C. Cao, X. Chen, and G. Xu. (May 2018). "Multimodal emotion recognition for one-minute-gradual emotion challenge." [Online]. Available: https://arxiv.org/abs/1805.01060
[45] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 2672–2680.
[46] J. F. Nash, Jr., "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci. USA*, vol. 36, no. 1, pp. 48–49, 1950.
[47] M. Mirza and S. Osindero. (Nov. 2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784
[48] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. (Sep. 2016). "Neural photo editing with introspective adversarial networks." [Online]. Available: https://arxiv.org/abs/1609.07093
[49] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 263–270.
[50] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proc. IEEE 13th Conf. Int. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 294–301.
[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[52] K. Pu, Z. Lian, and Z. Liu, "Multiple objects tracking based on multiple information integration," in *Proc. IEEE Int. Conf. Prog. Inform. Comput. (PIC)*, Dec. 2017, pp. 205–208.
[53] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.
[54] S. Ioffe and C. Szegedy. (Feb. 2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[55] M. Arjovsky and L. Bottou. (Jan. 2017). "Towards principled methods for training generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1701.04862
[56] M. Arjovsky, S. Chintala, and L. Bottou. (Jan. 2017). "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875
[57] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.
[58] J. Chen, J. Konrad, and P. Ishwar. (Mar. 2018). "VGAN-based image representation learning for privacy-preserving facial expression recognition." [Online]. Available: https://arxiv.org/abs/1803.07100
[59] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2242–2251.
[60] A. Mollahosseini, B. Hasani, and M. H. Mahoor. "AffectNet: A database for facial expression, valence, and arousal computing in the wild." [Online]. Available: https://arxiv.org/abs/1708.03985
[61] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 2584–2593.
[62] 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.
[63] P. Adam *et al. Pytorch*. Accessed: Mar. 5, 2018. [Online]. Available: https://github.com/pytorch/pytorch
[64] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie, "An occam's razor view on learning audiovisual emotion recognition with small training sets," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Boulder, CO, USA, 2018, pp. 589–593.
[65] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1967–1971.
[66] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *Proc. 25th Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1957–1961.
[67] F. Shen, J. Liu, and P. Wu, "Double complete D-LBP with extreme learning machine auto-encoder and cascade forest for facial expression analysis," in *Proc. 25th Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1947–1951.
[68] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.
[69] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Proc. ICANN*. Cham, Switzerland: Springer, 2018, pp. 84–94.
[70] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[71] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Comput. Vis. ECCV*, Munich, Germany, 2018, pp. 227–243.
[72] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-Gated CNN for occlusion-aware facial expression recognition," presented at the ICPR, Beijing, China, Aug. 2018.
[73] A. Gurnani, K. Shah, V. Gajjar, V. Mavani, and Y. Khandhediya. (Mar. 2018). "SAF-BAGE: Salient approach for facial soft-biometric classification–age, gender, and facial expression." [Online]. Available: https://arxiv.org/abs/1803.05719
[74] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

JIA DENG received the B.S. degree from the College of Physics, Northeast Normal University, Changchun, China, in 2017. She is currently pursuing the M.S. degree with the School of Mechanical Engineering, Zhejiang University. Her research interests include biomedical signal processing and machine learning for emotional interaction, and fluid human-robot collaboration/human–robot interaction.

GAOYANG PANG received the B.S. degree from the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China, in 2017. He is currently pursuing the M.S. degree with the School of Mechanical Engineering, Zhejiang University. His research interests include flexible sensing electronics, wearable sensors using inkjet printing technology, and safe human–robot collaboration strategies.

ZHIYU ZHANG received the B.S. degree from the College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou, China, in 2018. He is currently pursuing the M.S. degree with the School of Mechanical Engineering, Zhejiang University. His research interests include health IOT, bio-patch design, and implementation based on a low-power system-on-chip.

ZHIBO PANG (M'13–SM'15) received the B.Eng. degree in electronic engineering from Zhejiang University, Hangzhou, China, in 2002, the M.B.A. degree in innovation and growth from the University of Turku, Turku, Finland, in 2012, and the Ph.D. degree in electronic and computer systems from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2013. He was a Co-Founder and a CTO of startups such as Ambigua Medito AB. He is currently a Principal Scientist on wireless communications with the ABB Corporate Research Sweden, Västerås, Sweden, leading research on digitalization solutions for smart buildings and homes, robotics and factories, and power electronics and power systems. He is also serving as an Adjunct Professor or similar roles at universities such as the Royal Institute of Technology (KTH), Tsinghua University, China, and the Beijing University of Posts and Telecommunications, China. He is a Senior Member of IEEE. He is a Co-Chair of TC in the Technical Committee on Industrial Informatics. He is an Associate Editor of the IEEE Transactions on Industrial Informatics, the IEEE Journal of Biomedical and Health Informatics, and the IEEE Reviews in Biomedical Engineering, a Guest Editor of the IEEE Access, an Editorial Board of the *Journal of Management Analytics* (Taylor & Francis), the *Journal of Industrial Information Integration* (Elsevier), and the *International Journal of Modeling, Simulation, and Scientific Computing* (WorldScientific).

HUAYONG YANG received the Ph.D. degree in philosophy from the University of Bath, in 1988. He joined the Department of Mechanical Engineering, Zhejiang University, as a Postdoctoral Researcher, in 1989. He is currently a Professor and the Director of the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University. His research interests include motion control and energy saving of mechatronic systems, the development of fluid power component and systems, and integration of electro-hydraulic systems and engineering application. He has 169 invention patents. He has co-authored three academic books, over 76 Science Citation Index papers, and 210 Engineering Index papers published.

GENG YANG received the B.Sc. and M.Sc. degrees from the College of Biomedical Engineering and Instrument Science, Zhejiang University (ZJU), Hangzhou, China, in 2003 and 2006, respectively, and the Ph.D. degree in electronic and computer systems from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2013. From 2013 to 2015, he was a Postdoctoral Researcher with the iPack VINN Excellence Center, School of Information and Communication Technology, Royal Institute of Technology (KTH). He is currently a Research Professor with the School of Mechanical Engineering, ZJU. He developed low power and low noise bio-electric SoC sensors for m-health. His research interests include flexible and stretchable electronics, mixed-mode IC design, low-power biomedical microsystems, wearable bio-devices, human–computer interface, human–robot interaction, intelligent sensors, and the Internet-of-Things for healthcare.