Received February 28, 2020, accepted March 6, 2020, date of publication March 11, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980060

A Facial Expression Recognition Method Using Deep Convolutional Neural Networks Based on Edge Computing

AN CHEN[1], HANG XING[2], AND FEIYU WANG[3]

[1]Department of Experiment Teaching, Guangdong University of Technology, Guangzhou 510006, China
[2]College of Engineering, South China Agricultural University, Guangzhou 510642, China
[3]School of Transportation Management, Xinjiang Vocational & Technical College of Communications, Urumchi 831401, China

Corresponding author: Hang Xing (hangsail@163.com)

This work was supported by the Science and Technology Program of Guangdong Province under Grant 2015A020209106.

ABSTRACT The imbalanced number and the high similarity of samples in expression database can lead to overfitting in facial recognition neural networks. To address this problem, based on edge computing, a facial expression recognition method using deep convolutional neural networks is proposed. In order to overcome the shortcoming that circular consensus adversarial network model can only be mapped one-to-one, we construct a constrained circular consensus generative adversarial network by adding class constraint information. Discriminators and classifiers in this network can share network parameters. In addition, for the problems of unstable training and easy to encounter model collapse in original GAN networks, this paper introduces gradient penalty rule into discriminator's loss function to achieve the normative constraint on gradient changes. Using this network not only generates sample data for a few classes in the training set of expression database, but also performs effective expression classification. Compared with other methods, the improved discriminative classifier network structure can enhance the diversity of samples and get a higher expression recognition rate. Even if other expression feature extraction methods are used, the higher recognition rate can still be obtained after using proposed data augmentation framework.

INDEX TERMS Facial expression recognition, generative adversarial network, deep learning, edge computing, class constraint information, gradient penalty rule.

I. INTRODUCTION

With the rapid development of intelligent information technology and the wide use of computers, people can turn the complicated work to computers, which not only changes the traditional way of life, but also provides great convenience for human beings. However, in today's diversified human-computer interaction needs, the traditional single and fixed input-output mode has been unable to meet the needs of real life and market applications. We hope that computers can understand their intentions more intelligently, so as to better serve us. Facial expressions are an essential way to express emotions in nonverbal communication. With the development of computers, facial expression recognition plays an important role in many applications, such as human-computer interaction and medical escort. It becomes current research hotspots in the fields of artificial intelligence, computer vision, and even the Internet of Things [1]–[3]. However, facial expression recognition is a complex task for computers. The process of recognition facial expressions can be divided into three steps: image preprocessing, feature extraction and facial expression classification. How to effectively extract the features of facial expression images is a critical step in facial expression recognition. Early facial expression recognition methods were developed to extract facial expression features manually by designing feature extraction algorithms. These methods include active appearance model algorithms for face modeling based on feature point location and extraction algorithms based on local features, such as Garbor wavelet, Weber Local Descriptor (WLD), Local Binary Pattern (LBP), multi-feature fusion, etc. These artificially designed feature extraction methods may lose some feature information of original image and are not robust to image scale and lighting conditions.

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

Compared to manual feature extraction methods, deep neural networks can learn features automatically and achieve a high recognition rate in facial expression recognition. In order to extract more facial expression features, the number of layers of neural networks is also increasing gradually. However, these networks tend to overfit with the deepening of networks and the increase of parameters. The smaller data set, the more severe overfitting phenomenon. Most of facial expression data sets have problems of insufficient data and high sample similarity. Besides, imbalanced samples can also lead to unsatisfactory neural network recognition [4], [5].

Data augmentation is an important means to resolve sample shortages and imbalances. Reference [6] applied traditional rotation and crop data augmentation methods to expand the training samples. Most of the images carry duplicate information, which is close to a simple copy of the sample. Besides, this is still far from the same number of different samples in amount of information, and they do not change the identity information of images. Therefore, the problem of large sample similarity remains unsolved. Different from using simple geometric transformations and crop for data augmentation, GANs introduce an adversarial loss function and learns the facial expression images with the same distribution as target datasets, which can solve the high similarity problem of generated samples. However, GANs networks map random vectors into the target dataset. This is often due to the lack of constraints, resulting in uneven quality.

For the uneven number of facial expression samples, such as relatively small amount of disgust and sad expression data, this paper introduces Cycle GAN into the facial expression data augmentation. This enables the mapping of neutral expressions to multi-category expressions. At the same time, Cycle GAN has a one-to-one mapping relationship. Therefore, when there is a one-to-many mapping relationship (such as a neutral expression to a variety of expressions such as happy, sad and surprised), the model needs to be trained multiple times, which brings a huge time cost. To address this problem, this paper improves Cycle GAN and further proposes a constrained circular consensus generative adversarial network for facial expression recognition. The network introduces class constraint conditions and gradient penalty rules, and implements one-to-many mapping transformation in one model. It reduces the model training overhead while obtaining higher quality generated images.

Compared with the circular consensus generative adversarial network, this network has three major improvements:

1) Add an auxiliary expression classifier based on the discriminator. The newly added discriminative classifiers are used to replace the two discriminators of circular consensus generative adversarial network. And it can judge the authenticity of input images and classify expressions.

2) Aiming at the problems of unstable training and easy to encounter model collapse in original GAN networks, this paper introduces gradient penalty rules into the loss function of discriminator, which can achieve the normative constraint on gradient changes.

II. GANS-BASED EXPRESSION RECOGNITION APPLICATION

Facial expression image editing is a special and important research topic. Due to human vision is sensitive to facial irregularities and deformation, it is not easy to edit realistic facial expression images. In this regard, GANs can edit facial expression images with high-quality detailed textures. Moreover, the expression recognition on the edited expression images can still achieve good effect.

Facial expression editing is a challenging task because it requires advanced semantic understanding of the input facial images. In traditional methods, either the paired training data is needed, or the synthesized face image resolution is very low. Reference [7] proposed Conditional Adversarial Auto Encoder (CAAE) to learn face manifolds, and then realized smooth age expression image regression on the face manifold. In CAAE, the face image is first mapped to a latent vector by convolutional encoder and the vector is projected onto a face manifold based on age by deconvolution generator. Latent vectors retain the subject's facial features, and age conditions control regression. Using adversarial learning on the encoder and generator makes generated images more realistic. Experimental results show that the framework has good performance and flexibility, the quality of generated images is high. Reference [8] proposed an Expression Generative Adversarial Networks (Expr GAN) based on CAAE, which can edit the facial expression intensity of real images. In addition to the encoder and decoder networks, Expr GAN also designed an expression intensity control network specifically for learning expression intensity of generated images. This novel network structure allows the intensity of generated expression images to be adjusted from low to high.

Reference [9] proposed a Conditional Difference Adversarial Autoencoder (CDAAE), which is a facial expression synthesis based on AU tags. Enter an unseen face image and use the target expression label or facial action unit (AU) label to generate the person's facial expression image. CDAAE adds a feedforward path to autoencoder structure and it connects the low-level features of encoder with the corresponding levels features of decoder. By learning to distinguish the differences between low-level image features of different facial expressions for the same person, the problem of changes due to different identities and changes due to different facial expressions can be solved. Experimental results show that CDAAE can more accurately save the facial expression information of unknown objects than the latest methods. However, the resolution of facial images generated by CDAAE is only 32 × 32 and facial images with AU labels not be well evaluated quantitatively. Reference [10] combined the geometric model of 3DMM [11] with generative model. The former can separate expression attributes from other facial attributes and generate 3DMM expression attributes based on target AU label. In this way, we can generate high-resolution facial expression images and the target expression label is determined by AU label. Reference [12] proposed a method for augmenting facial expression data based on Cycle GAN.

They used neutral expressions to convert to other expressions, and expand expression images with less data, such as disgust and sad expressions. Consequently, the classification accuracy is improved by 5% -10% after using data augmentation technology.

Image restoration is a traditional graphics problem. It refers to restoring the missing part of images based on the existing information in images, so that human eyes cannot distinguish which part is restored. Something about image restoration involves statistics, probabilistic models, etc. [13]–[15]. The application of image restoration in facial expression recognition is very common. During the process of identifying facial expressions, key parts of facial images that may be identified may be occluded. For example, some facial images wear sunglasses and some wear scarves, which will block eyes or mouth and the effect of these occlusions on expression recognition is still significant. Therefore, the occlusion part can be restored by a design algorithm and then facial recognition is performed. The traditional methods are to restore an image by copying pixels from the original image or by copying patches from an image library, while GANs provide a new method for image restoration.

Reference [16] proposed a network structure of context encoder that is the first image restoration method based on GANs. The network is based on an encoder-decoder architecture and the network input is 128 * 128 images with missing blocks. The output is 64 * 64 missing content (when the missing block is in the middle of original image) or 128 * 128 full restore image (when the missing block is anywhere in original image). Besides, the objective function includes adversarial loss and content loss. Experimental results show that the restoration effect is better when missing block is in the middle of original image. Reference [17] proposed a method of semantic image restoration based on GANs iteration, which pre-trains GANs, and its generator maps the hidden variable z into an image. Enter an image 0x with some missing information, and encode 0x into *z by minimizing the objective function. Among them, the objective function includes an adversarial loss function and a content loss function. The content loss function calculates the weighted L1 pixel distance between generated image (*Gz) and the undamaged area on 0x. The pixel values near missing information area have a higher weight. Finally, the objective function is optimized by back propagation iteration. Reference [18] considered that GANs-based image restoration models are susceptible to the initial solution of non-convex optimization criteria and built an end-to-end trainable parametric network. They started with a good initial solution and get a more realistic image reconstruction with significant optimization speed and learned to use a recurrent neural network to optimize the time window of the initial solution. In the iterative optimization process, a time smoothness loss is applied to respect the redundancy of sequence time dimension. Experimental results show that this method is significantly better than other methods in image reconstruction quality. Reference [19] designed a facial image generative network based on Wasserstein GANs, which can generate context-complete complementary images and expression recognition networks for the occlusion areas in images. It extracts expression features and infers expression categories, and achieves a high recognition effect on CK+ database. All of these mechanisms have achieved good recognition results, but the quality of images generated by GANs network is uneven due to the lack of constraints. Moreover, it can't realize flexible expression mapping in the case of unbalanced number of facial expression samples.

Reference [20] proposed a facial restoration algorithm based on depth generative model. Different from the completion of well-designed background restoration studies, the facial restoration task is more challenging because it usually requires the generation of semantically new pixels for key missing parts of a large number of appearance changes, such as eyes and mouth. In reference [21], a novel cascaded backbone-branches fully convolutional neural network (BB-FCN) is proposed, which is used to locate face markers quickly and accurately in unconstrained and chaotic environment. BB-FCN does not need any preprocessing, and generates facial landmark response map directly from the original image. BB-FCN follows a cascade pipeline from coarse to fine, which is composed of a backbone network and a branch network, which is used to roughly detect the location of all facial marker points, and provide a branch network for each type of marker point detected to further refine its location. These mechanisms have achieved good recognition results. However, these mechanisms need multiple training models, which brings huge time cost.

In addition to the above two GANs-based expression recognition applications, there are other application methods. However, the applications of GANs in expression recognition are mostly used for data augmentation.

III. THE PRINCIPLE OF GAN

The GAN model contains two networks: generator G and discriminator D. the basic structure and calculation flow are shown in Figure 1.

Generator G has a nonlinear and differentiable deep neural network structure, which can be trained in the end-to-end mode. The objective function of generator G is to minimize the log likelihood function so that the data distribution of $G(z)$ is close to the data distribution of training sample x. The maximum log likelihood function is used to determine whether the data is from the training sample or from the generator. In short, the input of generator G is a random coding vector, and the output of generator G is a complex sample; the input of discriminator D is a complex sample, and the output of discriminator D is a probability, which indicates whether the sample of input discriminator D is a real sample or a false sample generated by generator. The purpose of the discriminator is to distinguish whether the input samples come from the real samples or the samples generated by the generator, while the purpose of the generator is to make the discriminator unable to distinguish whether

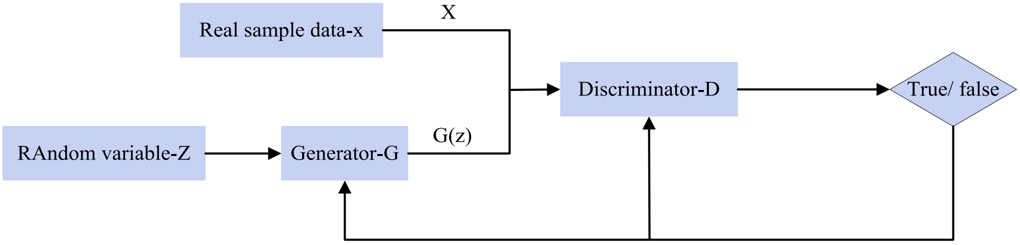


FIGURE 1. Basic structure and calculation flow chart of GAN.

the input samples come from the real samples or the samples generated by the generator. From this point of view, the goal of the generator and the discriminator is the opposite, and the two are antagonistic. Generator G and discriminator D use min max rule to train alternately. Through continuous iteration, update the parameters of generator G and discriminator D until the balance is reached. At this time, there is no difference between the data distribution of $G(z)$ and the real sample x. at the same time, discriminator D can not distinguish whether the data is from the real data or the output data of generator G.

The characteristics of the generated countermeasure network are as follows:

(1) The complexity of generating data is linearly related to the dimension. If we want to generate a larger image, we will not see the exponential increase of parameter calculation as the traditional generation model does;

(2) There are few prior hypotheses, and GAN does not make any display parameter distribution hypothesis for the data, that is, it does not need to know or assume the distribution of training data in advance, and only requires that the discriminant network and the generated network be differentiable;

(3) It can generate higher quality images, and the generation process does not depend on the form of Markov chain, and the confrontation between the generation network and the discrimination network can be realized through the back propagation algorithm; it does not need to gradually generate images in the form of pixel points like PixelRNN, but takes direct generation, and the generation speed is faster than other methods;

Generally speaking, the traditional discriminant model is an optimization function, and there must be an optimal solution for a convex optimization problem. However, GAN needs to take into account both the discriminant network and the generated network training situation. It is difficult to find out the Nash equilibrium point between the two networks, and it is difficult to determine the position of this equilibrium point through a simple gradient descent algorithm.

GANs is a milestone for the development of generation model. As a new generation method, GANs does not need to define the probability distribution of the generation model in advance. Instead, it can generate artificial samples that match the input samples through the counter learning of generators and discriminators, which can effectively solve the problem of high-dimensional data generation. The network structure of generator and discriminator has no limitation on the generation dimension, and widens the scope of generating samples. Compared with other generation models, GANs has the following advantages:

(1) GANs can sample the generated samples in parallel. The generator of GANs is a simple feedforward network which maps the hidden variable Z to the real sample x. it can generate data at one time, thus greatly speeding up the sampling speed.

(2) In the training process of GANs, there is no need to make all kinds of approximate reasoning, to use the inefficient Markov chain method, and to calculate the complex lower bound of variation, which greatly improves the training difficulty and training efficiency of the generation model.

(3) The training method of generator and discriminator increases the diversity of generating samples. Compared with other generation model experimental results, GANs can generate higher quality and clearer samples, which provides a possible solution for generating meaningful samples for human beings.

(4) GANs not only makes a great contribution to the generation model, but also provides inspiration in unsupervised learning. In semi supervised learning, we can first use unlabeled samples to train GANs, and then use a small part of labeled data to train discriminators for traditional classification and regression tasks.

IV. PROPOSED IMPROVED CYCLEGAN

The proposed improved CycleGAN achieves image style conversion from source domain to target domain. For the expression recognition task, the data augmentation mainly focuses on the conversion of neutral expressions to multiple expressions (anger, disgust, fear, sadness, surprise and joy). It uses circular consensus generative adversarial network to train multiple models for transformation. This section constructs a structure that constrained generates adversarial networks based on constrained cycles, as shown in Fig. 2. Add class constraint information to CycleGAN to achieve multi-category style conversion in one model. 2) Add an auxiliary expression classifier C to the discriminator. The newly added discriminative classifiers are used to replace the two discriminators of circular consensus generative adversarial

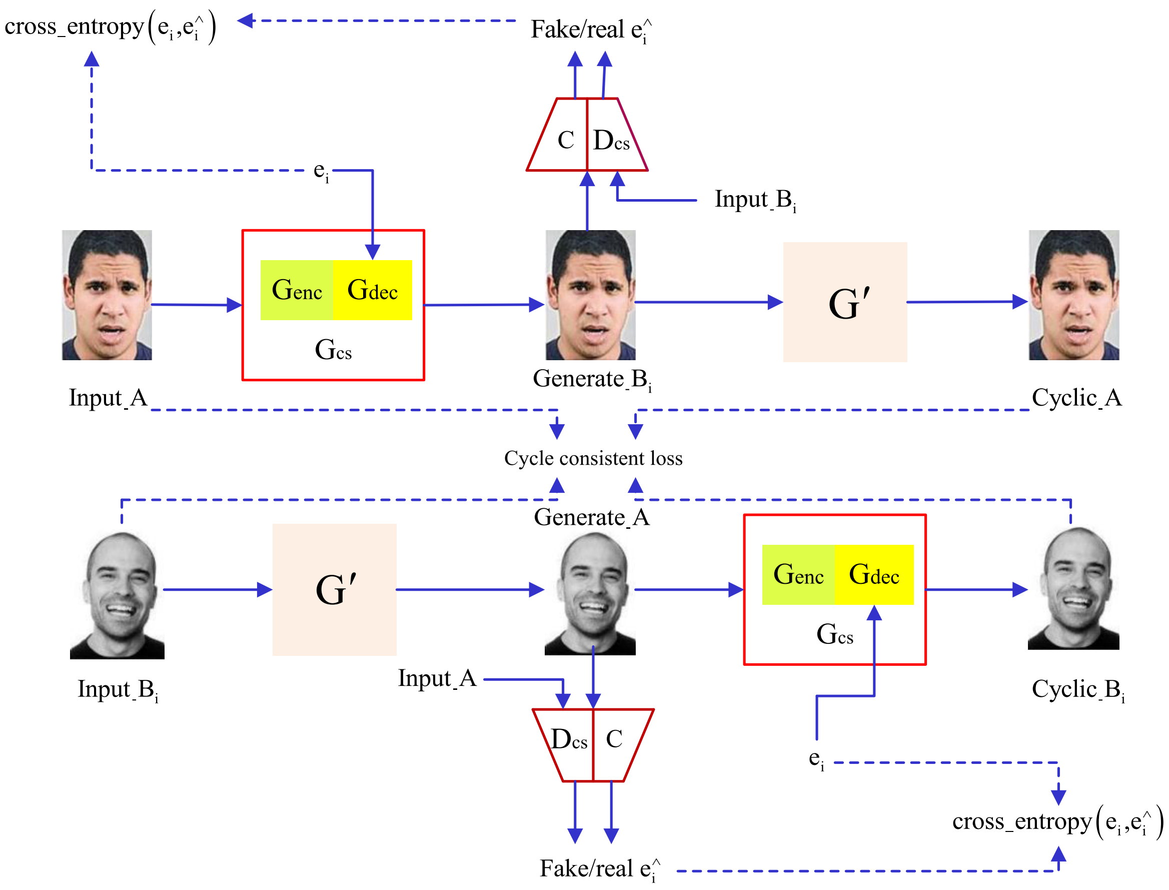


FIGURE 2. Structure of the generator.

network, which can judge the authenticity of input images and classify expressions.

In Fig. 2, the sample training set is divided into two parts, neutral expressions and other classified expressions. Neutral expressions are represented by source domain, and other classified expressions are represented by target domain and together form a mapping network from neutral expressions to other classified expressions, where is responsible for encoding the original picture. We concatenating it with the target domain class constraint information (such as: happy) after implicit coding, and then decoding. Thus, the target domain category expression image is obtained. The discrimination classifier is responsible for classifying and authenticating generated images, which is a many-to-one mapping network of multi-type expressions to neutral expressions. The loss function of network has an additional classification loss in addition to the adversarial loss and the cycle consistent loss:

A. ADVERSARIAL LOSS L_{adv}

Promote generator G_{cs} to transform the neutral expression pictures of source domain A into pictures that more closely approximate true multi-type expression distribution from multi-target domain $B_{i=1}^{S}$. Here S represents the total number of expression categories.

$$L_{LSGAN}\left(G_{cs}, D_{cs}, A, B_i\right) = E_{b\sim P_{data}(B_i)}\left[\left(D_{cs}(b)-1\right)^2\right] + E_{a\sim P_{data}(A)}\left[\left(D_{cs}\left(G_{cs}(a, e_i)\right)\right)^2\right] \tag{1}$$

Equation (1) represents the loss function to be optimized by mapping $A \rightarrow B_{i=1}^{S}$, and e_i represents category information.

$$L_{LSGAN}\left(G', D_{cs}, A, B_i\right) = E_{a\sim P_{data}(A)}\left[\left(D_{cs}(a) - 1\right)^2\right] + E_{b\sim P_{data}(B_i)}\left[\left(D_{cs}\left(G'(b)\right)\right)^2\right] \tag{2}$$

Equation (2) represents the loss function to be optimized by mapping $B_{i=1}^{S} \rightarrow A$.

$$L_{adv} = L_{LSGAN}\left(G_{cs}, D_{cs}, A, B_i\right) + L_{LSGAN}\left(G', D_{cs}, A, B_i\right) \tag{3}$$

Equation (3) represents the overall adversarial loss function.

B. CYCLE CONSISTENT LOSS

$$L_{cyc} = E_{a\sim P_{data}(A)}\left[\left\|G'\left(G_{cs}(a, e_i)\right) - a\right\|_1\right] + E_{b\sim P_{data}(B_i)}\left[\left\|G_{cs}\left(G'(b), e_i\right) - b\right\|_1\right] \tag{4}$$

C. CLASSIFICATION LOSS

As mentioned in original GAN network, the training is unstable and the model collapses easily. Most researchers in this problem area have proposed that this is caused by trying to minimize a strong divergence in network training. To solve this problem, we introduce gradient penalty rules into the discriminator's loss function which can regulate the gradient change. The gradient penalty used is shown in formula (5):

$$\lambda E_{x-\chi}\left[\|\nabla_x D(x)\|_p - 1\right]^2 \tag{5}$$

The introduced gradient penalty is not applied to the entire network area. It can apply to the real sample concentration area, the generated sample concentration area, and the area in middle of them. Thus, first randomly sample a pair of true and false samples and generate a random number in range [0-1] as follows:

$$x_r \sim P_r, \quad x_g \sim P_g, \ \varepsilon \sim Uniform[0, 1] \tag{6}$$

In formula (6), $x_r \sim P_r$ represents the area sampling of real sample concentration, and $x_g \sim P_g$ represents the area sampling of generated sample concentration. The ε value is a random number in the interval [0,1]. Then perform random interpolation sampling on the line between x_r and x_g:

$$\widehat{x} = \varepsilon x_r + (1-\varepsilon)\, x_g \tag{7}$$

The distribution satisfied by $\widehat{x}$ obtained by sampling according to above process is $P_{\widehat{x}}$, and the discriminator gradient punishes loss:

$$\lambda E_{x-\chi}\left[\|\nabla_x D(x)\|_p - 1\right]^2 \tag{8}$$

The generative path is mainly training generator G and discriminator D. The input to generator is not only random noise z, and a given view angle label v. The purpose of generator G is to generate a new image $G(v, z)$ in view v. The role of discriminator D is that the view distinguishes real data x from generated data $G(v, z)$. The loss function of generator G is shown in formula (9):

$$L_{G_{vzx}} = \underset{z\sim P_z}{E}\left[D_s\left(G(v,z)\right)\right] + \lambda_3 \underset{z\sim P_z}{E}\left[P\left(D_v\left(G(v,z)\right) = v\right)\right] \tag{9}$$

The discriminator D removes the input of view angle label and increases the output of classification angle and score. So the loss function of discriminator D is as shown in formula (10), where the third term in formula is a gradient loss function, and the fourth term is a cross-entropy loss function using ACGAN.

$$\begin{aligned} L_{D_{xvs}} = {} & \underset{z\sim P_z}{E}\left[D_s\left(G(v,z)\right)\right] - \underset{x\sim P_x}{E}\left[D_s(x)\right] \\ & + \lambda_1 \underset{\widehat{x}\sim P_x}{E}\left[\left(\left\|\nabla_{\widehat{x}} D(\hat{x})\right\|_2 - 1\right)^2\right] \\ & - \lambda_2 \underset{\widehat{x}\sim P_x}{E}\left[P\left(D_v(x) = v\right)\right] \end{aligned} \tag{10}$$

In the reconstruction path, encoder E and decoder D are mainly trained, encoder E attempts to reconstruct the training samples. The cross-reconstruction method is used in encoder E to reconstruct angle information from identity information to ensure that the images of multiple views have the same identity information. Specifically, samples (x_i, x_j) with the same identity but different angles are reconstructed from x_i to x_j. x_i is used as the input of encoder E and outputs a view estimate $\bar{v}$ and a representation $\bar{z}$ retained by identity information flag, that is, $(\bar{v}, \bar{z}) = (E_v(x_i), E_z(x_i)) = E(x_i)$. The resulting $\bar{z}$ and v_j are input into generator G together. Guided by angle v_j, G generates the corresponding $\hat{x}_j$. At this point $\hat{x}_j$ was refactored from x_i. Finally, the discriminator D tries to distinguish real x_j from generated $\hat{x}_j$, and obtains the corresponding score and angle information. In this network, the loss function of encoder E is shown in formula (11):

$$L_E = \underset{x_i,x_j\sim P_x}{E}\begin{bmatrix} D_s\left(\hat{x}_j\right) + \\ \lambda_3 P\left(D_v\left(\hat{x}_j\right) v_j\right) \\ -\lambda_4 L_1\left(\hat{x}_j, x_j\right) - \\ \lambda_5 L_v\left(E_v(x_i), v_i\right) \end{bmatrix} \tag{11}$$

Limiting $\hat{x}_j$ with E loss is x_j reconstructed from x_i. L_v loss is the cross-entropy loss between estimated view and real view. The loss function of discriminator D is:

$$\begin{aligned} L_{D_{xvs}} = {} & \underset{x_i,x_j\sim P_x}{E}\left[D_s\left(\hat{x}_j\right) - D_s\left(\hat{x}_j\right)\right] \\ & + \lambda_1 \underset{\widehat{x}_i\sim P_{\widehat{x}}}{E}\left[\left(\left\|\nabla_{\widehat{x}} D(\hat{x})\right\|_2 - 1\right)^2\right] \\ & - \lambda_2 \underset{x_j\sim P_x}{E}\left[P\left(D_v(x_i) = v_i\right)\right] \end{aligned} \tag{12}$$

Combining the above three parts of loss function, for the generators G and G', the loss function that needs to be optimized is:

$$\min_{G_{cs},G'} L_{G_{cs},G'} = \lambda_1 L_{cyc} + \lambda_2 L_{G_{vzx}} - L_{adv} \tag{13}$$

For discriminative classifier, the loss function that needs to be optimized is:

$$\min_{D_{cs},C} L_{D_{cs},C} = \lambda_3 L_{D_{xvs}} + L_{adv} \tag{14}$$

In the experiment, λ_1, λ_2, λ_3 in formula (13) and formula (14) are set to 100, 10, and 1, respectively.

V. FACIAL EXPRESSION RECOGNITION DIAGRAM BASED ON IMPROVED CYCLE GAN

On the basis of improving Cycle GAN, this paper proposes an efficient and secure facial expression recognition method based on the edge cloud framework combined with the improved Cycle GAN. The edge cloud computing framework is shown in Figure 3. In this system, the Internet of Things obtains facial expression signals from users through multi secret sharing technology, and then distributes them to different edge clouds to ensure the privacy of users.

Fig. 4 is a schematic diagram of the network structure of generator. The size of input facial expression images is 128 * 128. The coding structure of generator is a stack of five convolutional layers of same convolution kernel size and step

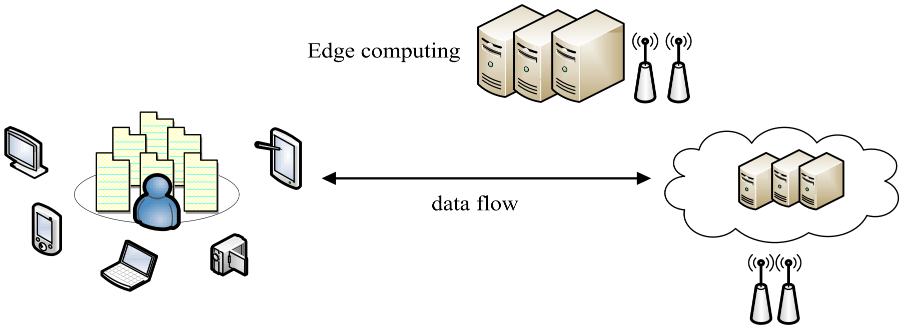


FIGURE 3. Structure of the edge computing.

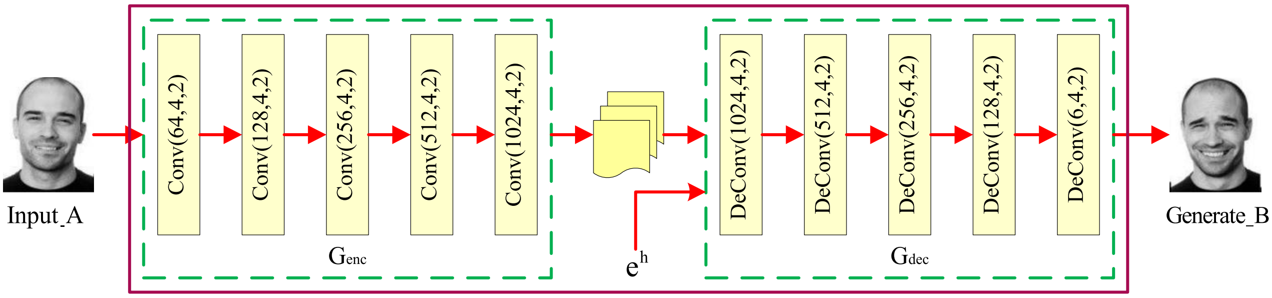


FIGURE 4. Structure of the generator.

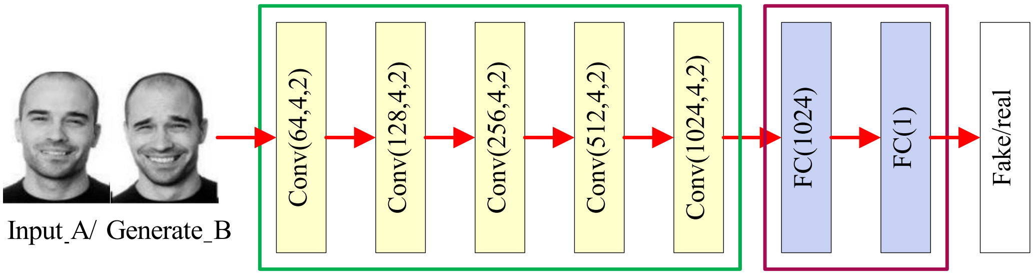


FIGURE 5. Structure of the discriminator.

size. First, the input image is dimensionally transformed into a (128, 128, 1) three-dimensional tensor. Using 5 convolutional layers with a size of 4 and a stride of 2, the number of convolutional output channels per layer is doubled. After five layers, the output is a (4,4,1024) three-dimensional tensor, which is equivalent to 1024 4 * 4 feature maps as the image encoding.

Fig. 5 is a schematic diagram of network structure of discriminator. The input sample image of (128, 128, 1) first passes through five convolutional layers with a size of 4 and a stride of 2. It then passes through two fully connected layers and outputs a one-dimensional discrimination result.

Fig. 6 is a schematic diagram of network structure of classifier. The network structure of classifier is similar to that of discriminator. It shares network parameters with discriminator, and the weights differ only in the last two fully connected layers.

VI. EXPERIMENT

A. EXPERIMENTAL RESULTS AND ANALYSIS OF JAFFE DATABASE

In this section, experiments are performed on JAFFE dataset [22], CK+ and FER2013 dataset [23], [24]. The experimental hardware environment is Intel (R) Core (TM) i7-7700K CPU @ 4.20 GHz processor, 16GB running memory (RAM), NVIDIA Geforce GTX 1080Ti GPU. The deep learning framework is Tensor Flow.

JAFFE dataset contains 213 images of 7 facial expressions (6 basic facial expressions +1 neutral) composed of 10 Japanese female models. Each image is rated as 6 emotion adjectives by 60 Japanese subjects. Contains seven types of facial expression grayscale images of anger, disgust, fear, neutrality, happiness, sadness and surprise. Since the size of output images generated by adversarial network is 128 * 128, when using generated

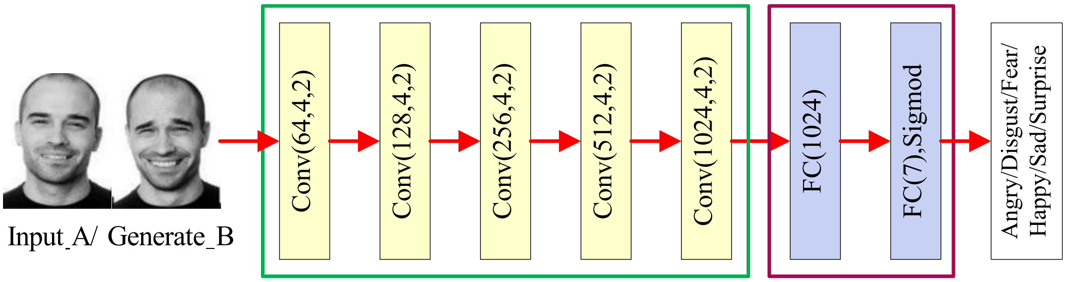


FIGURE 6. Structure of the classifier.

samples, this section uniformly processes the image samples to 48 * 48.

The amount of disgusted and surprised expressions in JAFFE database is far less than happy, sad and angry. This also stems from the biased objective fact of human emotion expression. Besides, the severely uneven distribution of data samples is an important reason limiting the increase in facial expression recognition rate. In addition to happy and angry expressions, neutral expressions also accounted for a larger proportion of experimental data. Mapping neutral expressions to aversions can alleviate sample imbalances. Therefore, the experiment uses the proposed improved Cycle GAN to augment JAFFE database.

Experiments use the structures of Fig. 4, 5, and 6 as generators, discriminators and classifiers of adversarial network respectively. Similarly, the model structure at expression recognition stage is the same as the classifier of Fig. 6. At the same time, the parameters are initialized randomly. In order to prevent overfitting during training, a Dropout layer with probability p = 0.9 is added after the first fully connected layer of discriminator and classifier. In addition, the parameters of all convolutional layers and all fully connected layers of discriminator are L2 regularized. Each layer of generator and discriminator uses Batch Normalization. Batch normalize the input of hidden layer to prevent the gradient from disappearing during training.

The model optimizer uses Adam, the learning rate is set to 0.0001, the momentum is 0.5 and the batch size is fixed to 32.

Adopt circular consensus generative adversarial network and constrained circular consensus generative adversarial network respectively. After 200,000 iterations, a sample comparison chart is generated. Since circular consensus generative adversarial network cannot achieve a one-to-many mapping relationship, it is necessary to implement a neutral expression to multiple expression mapping. Multiple models need to be trained, and the entire training takes about 4-5 times as much as the method proposed in this paper. In addition, due to the proposed improved Cycle GAN, classification loss is introduced. Following the attack, generated samples are more natural in terms of emotional expression.

In order to test the reliability of samples generated by network and the gain effect on facial expression recognition in this chapter. The neutral expression is used as the source domain and remaining expression classes are used as target domain. Consistently generative adversarial network samples through a constrained loop. Since the number of aversive expressions in JAFFE dataset is relatively small, more aversive expressions are generated to enhance the original data set, and a small number of samples are added to surprise expression category. Consequently, the final sample distribution is basically balanced. Some samples in test set were not augmented.

TABLE 1. Number distribution of seven expressions in FER2013 database.

Expression Classification	Training set	Test set
angry	3962	991
aversion	438	109
scared	4097	1024
happy	7191	1798
sad	4862	1215
surprised	3202	800
neutral	4598	1240

Fig. 7 shows the experimental results of expression classification on JAFFE dataset in this section. It can be seen from Fig. 7 that after data augmentation on original data set, not only the recognition rate of aversion expressions is improved, but also the recognition rate of other expressions also is improved. This is because as the number of training images increases, the differences between expressions increase. The more features we can get during training, the lower false recognition rate. The average recognition rate is improved accordingly.

B. EXPERIMENTAL RESULTS AND ANALYSIS OF FER2013 DATASET

In order to further verify the effectiveness of the method proposed in this chapter, we have carried out experiments on the FER2013 database. FER2013 facial expression data set consists of 35886 facial expression pictures, including 28708 training pictures, 3589 public test pictures and 3589 private test pictures. Each picture is composed of gray-scale images with fixed size of 48 × 48. There are 7 expressions, corresponding to digital labels 0-6. The expression number distribution of FER2013 database is shown in Table 1.

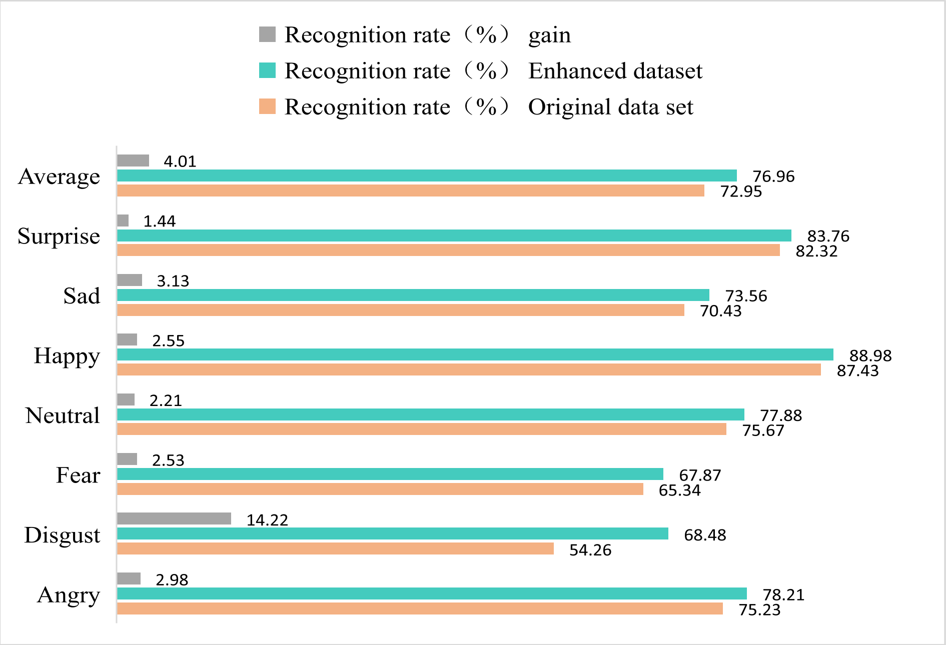


FIGURE 7. Recognition results on JAFFE dataset.

TABLE 2. Number distribution of seven expressions in FER2013 database after data augmentation.

Expression Classification	Training set	Test set
angry	3962	991
aversion	438(+4000)	109
scared	4097	1024
happy	7191	1798
sad	4862	1215
surprised	3202(+800)	800
neutral	4598	1240

TABLE 3. Recognition results on FER2013 database.

Expression Classification	RECOGNITION RATE (%)		
	Raw data	Enhanced dataset	gain
angry	75.23	77.43	2.20
aversion	56.87	73.98	17.12
scared	65.32	69.56	4.24
happy	87.43	88.67	1.24
sad	73.21	77.45	4.24
surprised	82.48	84.62	2.14
neutral	75.45	78.55	3.10
Average	73.71	78.61	4.90

TABLE 4. Number distribution of seven expressions in CK+ database.

Expression Classification	Training set	Test set
angry	113	28
aversion	131	37
scared	71	16
happy	183	45
sad	84	21
surprised	195	39
neutral	235	59

TABLE 5. Number distribution of seven expressions in CK+ database after data augmentation.

Expression Classification	Training set	Test set
angry	113(+470)	28
aversion	131(+470)	37
scared	71(+470)	16
happy	183(+470)	45
sad	84(+470)	21
surprised	195(+470)	39
neutral	235(+470)	59

From table 1, we can see that there are many samples of each expression category in FER2013 database, and the data of disgust and surprise expressions in FER2013 database is far less than that of happiness, sadness and anger. This is also due to the biased objective facts of human emotional expression. The seriously unbalanced distribution of data samples is an important reason to limit the improvement of facial expression recognition rate. All expression classes are enhanced by using constraint cycle to generate consistent adversary network. Table 2 shows the expression number distribution of FER2013 library after adding disgusting expressions and the number of test sets and training sets for facial expression classification. Because the number of disgusting expressions in FER2013 data set is relatively small, more disgusting expressions are generated to enhance the original data set, and a small number of samples are added to surprise expression category. The final sample distribution is basically balanced. Some samples of the test set were not enhanced.

Table 3 shows the experimental results of expression classification on FER2013 dataset. It can be seen from table 3 that

TABLE 6. Recognition results of different methods on CK+ database.

Method	Feature	Recognition rate before enhancement (%)	Recognition rate after enhancement (%)	Gain (%)
Reference [19]	Wasserstein GANs	94.35	95.56	1.21
Reference [25]	LeNet-5 cross connected	84.34	86.45	2.11
Reference [26]	combine IEw and ATRLBP operator	97.34	98.35	1.01
Proposed method	Cycle GAN + class constraint condition + gradient penalty rule	97.23	98.46	1.23

after data enhancement of the original data set, not only the recognition rate of disgusting expressions has been improved, but also the recognition rate of other expressions has been improved. This is because with the increase of the number of training images, more differences between expressions will be brought, and the more features can be obtained during training, the lower the error recognition rate and the corresponding increase of the average recognition rate.

C. EXPERIMENTAL RESULTS AND ANALYSIS OF CK+ DATASET

In order to further verify effectiveness of the method proposed in this paper, we conducted experiments on CK+ database. The CK+ database is extended from Cohn Kanade dataset and released in 2010. This database is much larger than Jaffe. It can also be obtained free of charge, including the label of expression and the label of action units. CK+ database includes 123 subjects and 593 image sequences. The last frame of each image sequence has the label of action units. Among these 593 image sequences, 327 have the label of emotion. This database is a popular database in facial expression recognition. Many articles will use this data for testing. In addition to the seven expressions in CK+ database, contempt expressions are also included. In order to be consistent with JAFFE database, we removed contempt expressions from the database in our experiments. A total of 1,249 facial expressions were extracted from the remaining seven types. The distribution of the number of expressions in CK+ library in the experiment is shown in Table 4.

From Table 4, we can see that there are fewer samples of each expression category in CK+ database. Adopt constrained loops to constrained generative adversarial networks for data augmentation for all expression classes. Before that, the mirror image flip operation was performed on neutral expression images to obtain 470 neutral expression images. The expression distribution of CK+ library after data augmentation is shown in Table 5.

Adopt circular consensus generative adversarial network and constrained circular consensus generative adversarial network respectively. After 100,000 iterations, a sample comparison chart is generated. In order to verify the effectiveness of network proposed for expression classifier training in this paper, the recognition rate on CK+ dataset is compared with the methods in other literatures. Table 6 shows comparison results of the recognition rates of different recognition methods on CK+ dataset. Reference [19] considered the individual differences of facial expressions during facial expression recognition and increased the amount of feature extraction. Reference [25] made up for the lack of training samples by combining low-level features with high-level features. Reference [26] considers the problem of low recognition rate due to the influence of various factors such as lighting, pose, expression, occlusion and noise on face recognition. They proposed a face recognition method (IE (w) ATR-LBP) combining weighted information entropy (IEw) with adaptive threshold circular local binary pattern (ATRLBP) operator.

The training classification network proposed in this paper based on constrained circular consensus generative adversarial network benefits from optimizing both the adversarial loss and the classification loss. The recognition rate on CK+ is 97.23%, which is higher than other methods. In addition, based on the reference [19], [25], [26], the methods proposed in them are trained using data-enhanced data. It can be seen from Table 3 that enhanced training set improves the recognition rate of all methods on test set. It brings an average gain of not less than 1.01%. For reference [19], the gain is less obvious. This may be because it suppresses identity information in loss function, while other types of facial expression samples mapped from neutral expressions retain the original identity information. For reference [25], since it is enhanced with geometric data such as rotation and clipping, the gain brought by it is obvious.

VII. CONCLUSION

Facial expression recognition is an important research content of computer vision and artificial intelligence, which is widely used in security, automatic driving, business and other aspects. Facial expression database is the data base of facial expression recognition, which plays an important role in the development of facial expression recognition technology. In this paper, the shortcomings of traditional data enhancement methods are analyzed and summarized. Aiming at the problem of class imbalance in the existing facial expression database, the paper improves Cycle GAN, proposes a method of facial expression recognition based on constraint cycle consistent generation to resist network, and introduces class constraint condition and gradient penalty rule. The experimental results show that the improved generation model can better learn the detailed texture information of the face image, and the quality of the generated image is high. The improved discriminator network has better classification and recognition effect on the enhanced face expression image recognition.

This paper studies facial expression recognition and expression image data enhancement. Although some achievements have been made, there are still some deficiencies, which need further research and improvement. First of all, the expression recognition and data enhancement in this paper are based on the static image, while the emotional changes in real life are of a certain timing, and the static image can only reflect the expression state of a person at a certain time. The next work will focus on the data enhancement of video sequence. Secondly, in the process of data enhancement, neutral expression image is used as the source domain and other expression images as the target domain, but the expression state of human in the real scene can be transformed at will. How to enhance the data without limiting the expression state of the input image is also a direction that can be improved in the future.

REFERENCES

[1] M. Zhang, Q. Fu, Y.-H. Chen, and X. Fu, "Emotional context modulates micro-expression processing as reflected in event-related potentials," *PsyCh J.*, vol. 7, no. 1, pp. 13–24, Mar. 2018.

[2] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2932058.

[3] H. Gao, "Transformation-based processing of typed resources for multimedia sources in the IoT environment," *Wireless Netw.*, to be published, doi: 10.1007/s11276-019-02200-6.

[4] X. Ma, H. Gao, H. Xu, and M. Bian, "An IoT-based task scheduling optimization scheme considering the deadline and cost-aware scientific workflow for cloud computing," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 249, Dec. 2019, doi: 10.1186/s13638-019-1557-3.

[5] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, "Context-aware QoS prediction with neural collaborative filtering for Internet-of-Things services," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2019.2956827.

[6] A. T. Lopes, E. de Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," in *Proc. 28th Conf. Graph., Patterns Images (SIBGRAPI)*, Salvador, Brazil, Aug. 2015, pp. 273–280.

[7] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. CVPR*, 2017, pp. 5810–5818.

[8] D. Hui, S. Kumar, and C. Rama, "ExprGAN: Facial expression editing with controllable expression intensity," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 6781–6788.

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

[10] Z. Liu, G. Song, J. Cai, T.-J. Cham, and J. Zhang, "Conditional adversarial synthesis of 3D facial action units," *Neurocomputing*, vol. 355, pp. 200–208, Aug. 2019, doi: 10.1016/j.neucom.2019.05.003.

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

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, Jan. 2012, vol. 25, no. 2, pp. 1097–1105.

[13] H. Gao, "Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data," *Intell. Automat. Soft Comput.*, vol. 25, no. 3, pp. 547–559, 2019.

[14] K. Xia, H. Yin, P. Qian, Y. Jiang, and S. Wang, "Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal CT images," *IEEE Access*, vol. 7, pp. 96349–96358, 2019.

[15] H. Gao, W. Huang, Y. Duan, X. Yang, and Q. Zou, "Research on cost-driven services composition in an uncertain environment," *J. Internet Technol.*, vol. 20, no. 3, pp. 755–769, 2019.

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

[17] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jul. 2017, pp. 5485–5493.

[18] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with preprocessing stage in automatic emotion recognition," *Procedia Comput. Sci.*, vol. 116, pp. 523–529, Oct. 2017.

[19] N. M. Yao, "Robust facial expression recognition with generative adversarial networks," *Acta Automatica Sinica*, vol. 44, no. 5, pp. 865–877, 2018.

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

[21] L. Liu, G. Li, Y. Yu, Q. Wang, L. Lin, and Y. Xie, "Facial landmark machines: A backbone-branches architecture with progressive representation learning," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2248–2262, Sep. 2019.

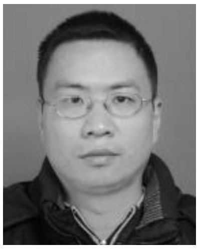
[22] A. V. Anusha, J. K. Jayasree, A. Bhaskar, and R. P. Aneesh, "Facial expression recognition and gender classification using facial patches," in *Proc. ComNet*, Jul. 2016, pp. 200–204, doi: 10.1109/CSN.2016.7824014.

[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPRW*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

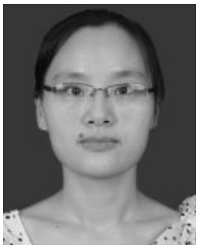
[24] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 558–565.

[25] Y. Li, X. Z. Liu, and M. Y. Jiang, "Facial expression recognition with cross-connect LeNet-5 network," *Acta Automatica Sinica*, vol. 44, no. 1, pp. 176–182, Jan. 2018.

[26] L. Ding, "Face recognition combining weighted information entropy with enhanced local binary pattern," *J. Comput. Appl.*, vol. 39, no. 8, pp. 2210–2216, 2019.



AN CHEN received the master's degree in control theory and control engineering from the Wuhan University of Technology, in 2006. He worked with the Guangdong University of Technology, where he is currently a Lecturer. His research interests include image recognition and analysis, and electronic science and technology.



HANG XING received the Ph.D. degree in agricultural electrification and automation from South China Agricultural University, in 2014. She worked with South China Agricultural University, where she is currently a Lecturer. Her research interests include image recognition and analysis, and control theory and engineering.

FEIYU WANG received the Master of Science degree in software engineering from the Beijing University of Technology, in 2009. She worked with the Xinjiang Vocational & Technical College of Communications, where she is currently an Associate Professor. Her research interests include intelligent algorithm, software engineering, and E-commerce.