IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 3, 2020, accepted August 31, 2020, date of publication September 3, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021531

LAUN Improved StarGAN for Facial Emotion Recognition

XIAOHUA WANG[1,2,3], JIANQIAO GONG[1,2,3], MIN HU[1,2], YU GU[1,2], (Senior Member, IEEE), AND FUJI REN[2,4], (Senior Member, IEEE)

[1]Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
[2]Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
[3]Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei 230601, China
[4]Graduate School of Advanced Technology & Science, University of Tokushima, Tokushima 7708502, Japan

Corresponding authors: Yu Gu (hfut_bruce@hfut.edu.cn) and Min Hu (jsjxhumin@hfut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672202, in part by the State Key Program of NSFC-Shenzhen Joint Foundation under Grant U1613217, and in part by the Fundamental Research Funds for the Central Universities of China under Grant PA2019GDPK0076.

ABSTRACT In the field of facial expression recognition, deep learning is extensively used. However, insufficient and unbalanced facial training data in available public databases is a major challenge for improving the expression recognition rate. Generative Adversarial Networks (GANs) can produce more one-to-one faces with different expressions, which can be used to enhance databases. StarGAN can perform one-to-many translations for multiple expressions. Compared with original GANs, StarGAN can increase the efficiency of sample generation. Nevertheless, there are some defects in essential areas of the generated face, such as the mouth and the fuzzy side face image generation. To address these limitations, we improved StarGAN to alleviate the defects of images generation by modifying the reconstruction loss and adding the Contextual loss. Meanwhile, we added the Attention U-Net to StarGAN's generator, replacing StarGAN's original generator. Therefore, we proposed the Contextual loss and Attention U-Net (LAUN) improved StarGAN. The U-shape structure and skip connection in Attention U-Net can effectively integrate the details and semantic features of images. The network's attention structure can pay attention to the essential areas of the human face. The experimental results demonstrate that the improved model can alleviate some flaws in the face generated by the original StarGAN. Therefore, it can generate person images with better quality with different poses and expressions. The experiments were conducted on the Karolinska Directed Emotional Faces database, and the accuracy of facial expression recognition is 95.97%, 2.19% higher than that by using StarGAN. Meanwhile, the experiments were carried out on the MMI Facial Expression Database, and the accuracy of expression is 98.30%, 1.21% higher than that by using StarGAN. Moreover, experiment results have better performance based on the LAUN improved StarGAN enhanced databases than those without enhancement.

INDEX TERMS Facial expression recognition, data enhancement, generative adversarial networks, self-attention.

I. INTRODUCTION

According to the research of psychologist Mehrabian, in the process of human communication, only 7% of information is transmitted through language. In contrast, the amount of information conveyed by the facial expression is as high as 55% [1]. Therefore, it is necessary to interpret facial expression information effectively. Affective computing proposed by Picard is one of the most important embodiments in human-computer interaction and artificial intelligence [2]. Facial expression recognition is an important means to realize affective computing. At present, facial expression recognition is a task that gets an extensive study and plays an integral part in digital entertainment, medical care, human-computer interactions, etc. Ekman, a famous American psychologist, and his team were the first to apply facial expression recognition to clinical cases [3]. In infant care, facial expression recognition could effectively and timely understand the state of babies [4].

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

Researchers are increasingly committed to giving computers the ability to perceive, recognize, and respond to human emotions. They are developing wearable computer systems and robots that can actively observe and give feedback [5].

Traditional expression recognition is divided into three stages: preprocessing, feature extraction, and facial expression classification. It is the key stage for facial expression recognition to extract the features of expression. Compared with traditional feature extraction algorithms, the deep learning method can achieve better performance. However, it is a challenge for facial expression recognition that deep neural networks need large-scale data because the scale of emotion datasets are limited [6]. Among existing methods, deep and complex network structures often cause overfitting. Therefore, data enhancement is an effective method to improve the performance of deep neural networks [7]. Traditional image enhancement methods are geometric transformation [8] and Color space transformation [9]. Simard *et al.* [10] proposed to use rotation, translation, and tilt of the original images to increase the number of samples. By combining these three spatial transformations, they got a great quantity of samples. Wang *et al.* [11] increased the number of samples by changing the brightness of original images, which reduced the influence of light on expression recognition to a certain extent. Hu *et al.* [12] proposed Fusion Features of Center-Symmetric Local Signal Magnitude Pattern for feature extraction. This method effectively improves the rate of expression recognition. Unlike simple transformation pixel and geometric transformation, Generative Adversarial Networks (GANs) can learn the features from the target dataset by introducing loss function. GANs have a generator and discriminator, and they are against each other to produce images [13]. Zhou *et al.* [14] proposed CycleGAN, which used the combination of adversarial loss and cyclic consistency loss to translate images. Luan *et al.* [15] proposed disentangled representation learning GANs for Generating face with different poses. GANs can be used in data enhancement because it can produce images with abundant features. Therefore, we used it for data enhancement of facial expressions.

StarGAN [16] is a model that can produce multiple domains of images from one image. GANs can get different expression facial images when inputting one facial image and retain identity information of the inputting image. Therefore, it is a proper way to extend the dataset. However, the images generated by StarGAN still have some flaws in some areas of the face, especially the face image with the 180-degree side face. To promote the quality of created pictures, we especially improved the StarGAN by changing the reconstruction loss and adding the Contextual loss. Moreover, we replaced StarGAN's generator with the Attention U-Net network to promote the model's performance. The experiments showed that Contextual loss and Attention U-Net (LAUN) improved StarGAN can generate higher quality facial images. In the existing facial expression recognition methods, many experiments are based on the frontal face, while the non-frontal face experiments are more challenging. There are some difficulties in data enhancement of non-frontal faces. One is that the rotation of head would cause part of the face to be occluded, and then the identity information of the face will be lost [17]. The other is that the shape of face texture presents nonlinear distortion with the change of posture, which makes confusion between different people [18]. In our experiments, we proved that the generated facial images still retain the person's identity as same as the original images. We also proved that the model can still produce facial images with different expressions when the original people of images have different pose, not only frontal facial images. After enhancing datasets with generated facial images, the generated pictures and the original images were inputted in VGG16 [19] for emotion classification. VGG16 is a convolutional neural network model proposed by simonyan and zisserman. Because of its excellent performance in classification, VGG16 is widely used in various classification tasks. In our model, the accuracy of expression recognition in VGG16 is improved over original pictures.

In brief, our contributions are as below.

1) We proposed a model basing on StarGAN, which is used to generate facial images with different emotions. Meanwhile, it can retain the identity of a person.
2) We changed the reconstruction loss and added the Contextual loss, which can promote the quality of generated facial images.
3) We changed StarGAN's generator and replaced it with the Attention U-Net network to enhance performance.
4) The experiments showed that the model is able to produce a human face with multiple emotions and different poses. Meanwhile, the experimental results proved that our improvement methods can alleviate some of the flaws in the face generated by the original StarGAN.

II. RELATED WORK

A. FACIAL EXPRESSION RECOGNITION

The facial expression has been widely adopted in various fields. In image retrieval, more and more applications implement retrieval through image semantic index [20]–[22]. Emotional semantics are included in image retrieval, and facial expression is the primary pattern manifestation for image semantic. Traditional facial expression recognition algorithms need to extract features of facial expression images manually. Dennis Gabor invented the Gabor wavelet [23], which is the application of Fourier transform in information theory. Gabor describes the relationship between pixels according to the filtered image results, which greatly reduces the influence of illumination. It is also a common feature extraction method. Ojala *et al.* [24] proposed the Local Binary Pattern (LBP), which is insensitive to noise. Simultaneously, the LBP is easy to calculate and has excellent performance, which makes subsequent researchers carry out a lot of improvement and optimization. The Histogram of Oriented Gradient [25] is based on the local feature descriptor. Because of its low computational complexity and strong feature description ability, it is widely used in facial expression feature extraction research.

Facial expression recognition is a complex task for intelligence computation [26]. In machine learning, many methods can be used to classify facial expression images. The AdaBoost algorithm trains several different weak classifiers for the training set and combines the weak classifiers into strong classifiers with strong classification ability [27]. Xing *et al.* [28] used the AdaBoost classifier to classify the texture features of facial expressions and achieved good experimental results. Prabhakar *et al.* [29] applied the Support Vector Machine (SVM) model to classify facial expressions. Compared with the AdaBoost method, it is proved that the SVM model has certain advantages in solving the classification problem of small samples.

In recent years, deep learning has become a popular method because of its good performance [30]–[32]. Compared to the manual feature extraction method, deep learning can automatically learn features and achieve a higher recognition rate in facial expression recognition. Yu *et al.* [33] used 9-layer Convolutional Neural Networks to carry out experiments on facial expression datasets, which greatly improves the recognition effect compared with traditional artificial features. Kuo *et al.* [34] proposed an expression recognition architecture based on image frames and image sequences, which reduced the number of network parameters. Meanwhile, a hybrid illumination enhancement scheme was proposed to alleviate the overfitting problem in the training process. Wang *et al.* [35] proposed the SelfCure Network, which suppresses the uncertainties efficiently and avoids model overfitting on uncertain facial images.

However, with the deepening of networks and the increasing of parameters, networks would appear overfitting phenomenon. Data enhancement is a significant way to solve database shortage and imbalance. Zhan *et al.* [36] presented a model for facial expression generation, relying on Condition Adversarial Autoencoder (CAAE) [37]. Ding *et al.* [38] proposed the ExperGAN, which can edit facial expressions and control the intensity of expression generation. Wang *et al.* [39] proposed Comp-GAN, which can transform facial expression and posture according to different input images. Bozorgtabar *et al.* [40] proposed a network named ExprADA, which can realize image to image conversion and achieve good performance in expression image generation. StarGAN has achieved excellent performance in face feature transformation [41]. Therefore, our model is based on StarGAN, and StarGAN is verified that it can retain the identity characteristics of the person in different postures to generate different expressions. However, we found that there are still some flaws in the generated face image by StarGAN. In order to further get higher quality generated pictures, we improved StarGAN's loss function and changed StarGAN's generator.

B. GENERATIVE ADVERSARIAL NETWORKS

GANs is extensively researched in recent years. The original GANs model consists of a generator and a discriminator. The training is carried out with a min-max two-player game between the two modules. The generator learns to generate images that are difficult to distinguish between real and fake for the discriminator, while the discriminator learns to distinguish between real and fake images. Conditional GANs [42] was to add conditional information based on GANs to control the generation of pictures. Conditional GANs enables the model to achieve supervised learning. The DCGAN [43] uses Convolutional Neural Networks to replace the multi-layer perceptron in discriminator and generator. Meanwhile, in order to make the whole network differentiable, the pooling layer in the network is removed. The full connection layer is displaced by the global pooling layer to reduce the computation. Karras *et al.* [44] proposed StyleGAN, an effective model to generate high-resolution images and transform style. Donahue *et al.* [45] proposed BigBiGAN, which can generate large-scale high-definition images. Meanwhile, the BigBiGAN can also be used in unsupervised learning. At present, GANs has made remarkable achievements in the field of image generation. GANs can effectively generate realistic images, making them widely used in image translation, data enhancement, style conversion, and other fields.

III. PROPOSED METHOD

A. STAR GENERATIVE ADVERSARIAL NETWORKS

StarGAN can translate an image to images with multiple domains. After inputting both one image and label, the generator learns to translate it into the corresponding domain flexibly. The structure of StarGAN is shown in Fig. 1.

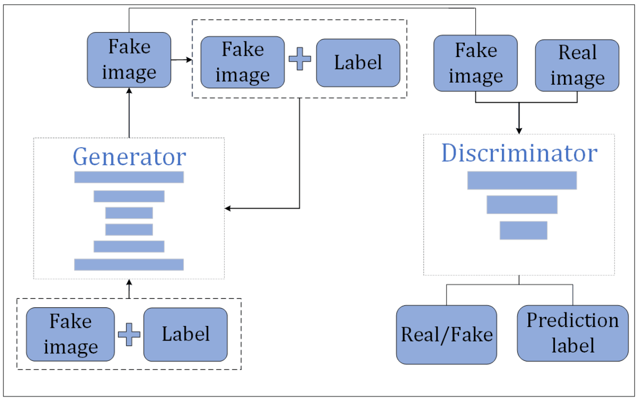


FIGURE 1. The structure of StarGAN.

1) ADVERSARIAL LOSS

The discriminator makes a distinction between real and fake images. Image x and target domain label c are inputted into the generator to generate the output image $G(x, c)$. A min-max objective function makes the fake pictures just like real.

$$L_{adv} = E_x[\log D_{src}(x)] + E_{x,c}[\log(1 - D_{src}(G(x, c)))] \quad (1)$$

The D_{src} is the result of the classifier to judge whether the picture is true or false. The generator expects the smaller the L_{adv}. The discriminator is the opposite.

2) CLASSIFICATION LOSS

The discriminator has two tasks. One is to distinguish the real or false images, and the other is to distinguish the category of the picture. The classification loss of real images trains the discriminator to make it find the real label of real images. The formula for classifying real pictures is as follow:

$$L^r_{cls} = E_{x,c'}[-\log D_{cls}(c'|x)] \tag{2}$$

The c' is the expression label of the input image x. The $D_{cls}(c'|x)$ is result of real images classification by the discriminator. Meanwhile, the generator tries to produce fake images that can be classified as the target label c. Specifically, the equation for classifying generated pictures is defined as:

$$L^f_{cls} = E_{x,c}[-\log D_{cls}(c|G(x,c))] \tag{3}$$

The $D_{cls}(c|G(x,c))$ is the result about discriminator distinguishing the generated picture. The discriminator should try to classify the real images correctly, so it should minimize L^r_{cls}. Furthermore, the generator tries to minimize L^f_{cls} for generating images, which can be categorized into the label c.

3) RECONSTRUCTION LOSS

The reconstruction loss can make reconstructive pictures similar to original pictures. The generator must preserve the represents of original pictures, and the generated pictures must retain the same represents. Therefore, reconstruction loss is as follow:

$$L_{rec} = E_{x,c,c'}[||x - G(G(x,c),c')||_1] \tag{4}$$

Image x and label c generate the target image $G(x,c)$ in the generator. Then $G(x,c)$ and original label c' use generator again to generate the reconstructed image $G(G(x,c),c')$. The L1 loss function is minimized by the generator.

4) FULL OBJECTIVE

The final training loss function has two parts. The loss function of the generator is as follows:

$$L_D = -L_{adv} + \lambda_{cls} L^r_{cls} \tag{5}$$

The loss function of the discriminator is as follows:

$$L_G = L_{adv} + \lambda_{cls} L^f_{cls} + \lambda_{rec} L_{rec} \tag{6}$$

λ_{cls}, λ_{rec} are the hyperparameter of the equilibrium loss.

B. CONTEXTUAL LOSS

Training Convolutional Neural Networks for image transformation relies on loss functions. Calculating the difference between the corresponding pixels is a common method to compare the different images. However, it is not a good way to deal with how to evaluate the similarity between images. When the original image and the generated image are aligned in space, the loss function to calculate the difference between pixels can achieve good results. However, if two images are not aligned in space, the image generation's effect using the traditional loss function is not ideal. Unlike the common methods, the Contextual loss [46] compares the images features to judge the similarity of images. Some deep neural networks can extract a series of high-dimensional features, which can be used to compare the similarity between images. VGG19 [19] is a classification Convolution Neural Network, which has the ability to extract image features. We use VGG19 as the feature extraction network. Cosine distance can calculate the similarity between different points. The Cosine distance between x_i and y_i is defined as follow:

$$d_{ij} = (1 - \frac{(x_i - u_y)\cdot(y_j - u_y)}{||x_i - u_y||_2||y_i - u_y||_2}) \tag{7}$$

where $u_y = \frac{1}{N}\sum_j y_j$. Then normalize the distance:

$$\tilde{d}_{ij} = \frac{d_{ij}}{\min_k d_{ik} + \varepsilon} \tag{8}$$

The ε is a tiny constant. The distance $\tilde{d}_{ij}$ is transforming into similarity calculation by power operation:

$$w_{ij} = e^{\frac{1-\tilde{d}_{ij}}{h}} \tag{9}$$

where $h>0$ is a bandwidth parameter. The contextual similarity between features is normalized similarity and scale-invariant:

$$CX_{ij} = \frac{w_{ij}}{\sum_k^N w_{ik}} \tag{10}$$

From the above formulas, we get the similarity between feature y_j and feature x_i. For each feature point y_j in the target image y, traverse each feature point x_i in the image x to find the most similar feature x_i with the feature y_j. Finally, the sum of the corresponding feature similarity on all y_j is the similarity of the two images. The sum of all y_j can be calculated to get the contextual similarity. The formula is as follow:

$$CX(x,y) = CX(X,Y) = \frac{1}{N}\sum_j \max_i CX_{ij} \tag{11}$$

x and y are images to be calculated, and X and Y are the characteristic graphs of the two images. x_i and y_i are the points in the two characteristic graphs, respectively.

Finally, we get the loss function. The features can be extracted from two images, and we can get the contextual similarity by calculating the above formulas. The contextual loss is as follow:

$$L_{CX}(x,y,l) = -\log(CX(\Phi^l(x), \Phi^l(y))) \tag{12}$$

$\Phi^l(x)$ and $\Phi^l(y)$ are the feature maps extracted by VGG19, and l is the number of layers in VGG19.

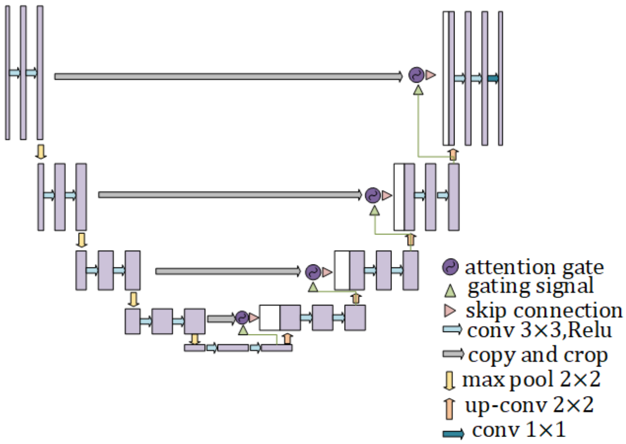


FIGURE 2. The network architecture of Attention U-Net.

C. ATTENTION U-NET

Attention U-Net [47] originates from the U-Net network [48], and attention mechanism is added to U-Net network. The model is used for image segmentation. It also uses attention models to restrain the uncorrelated areas and highlight the important features for specific tasks in the inputting image. Fig. 2 is the model structure of Attention U-Net. The network downsamples the image four times for feature extraction, and Upper sampling adopts the skip connection in the same stage. The structure of Attention Gate is shown in Fig. 3.

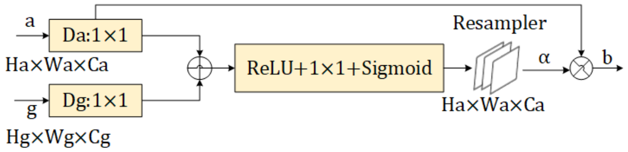


FIGURE 3. The structure of Attention Gate.

In Fig. 3, a is input features, and the α is attention coefficients computed by Attention Gate. Meanwhile, g is the gating signal gathered from a wide range, and it can determine focus regions for each pixel. Attention Gate's output is b, which is the element-wise multiplication for input feature-maps and attention coefficients. The a is a characteristic graph from the corresponding downsampling layer. The g is the feature image of the previous layer in the upsampling process. Da and Dg are convolution 1×1 kernels. Da convolutes a and Dg convolutes g. The convolution results are added, and the sum passes through ReLU, convolution, sigmoid, and resample. Finally, the result is α. α is the attention coefficient, and the size is consistent with a. The range of α is 0 to 1, which can make the value of the vital region in the feature map larger and the value of the unimportant region smaller. The output of the Attention Gate is obtained by multiplying g with the feature coefficient α. Attention Gate uses additive attention for fusion and calculates a single scalar attention value for each pixel vector. It is fused for the structure information of the lower sampling layer and the texture information of the current layer. And then, sigmoid normalization is used to obtain the regions with a strong correlation. The resulting α can identify significant image regions and prune feature responses to retain the activation associated with a specific task. α is multiplied with the current layer to emphasize the characteristics of the significant region in this layer. Attention Gate is added to the skip connection, and the output is added to the feature graph in upsampling. Therefore, one of the inputs in Attention Gate is the input in skip connection, which is the characteristic graph of downsampling in the corresponding layer. The other input of Attention Gate is the feature map of the previous layer in upsampling, as shown in Fig. 2.

Attention Gate learns to focus on the significant areas in inputting and contain contextual information passes through the Attention Gate. These Attention Gates generate soft regions that can get implicit information for highlight salient features [49]. Meanwhile, they do not bring a large number of computational costs, nor do they need as many model parameters. In brief, Attention U-Net is an improvement of the U-Net model, which promotes the sensitivity of the model to foreground pixels without sophisticated heuristics. The model extracts image features on multiple image scales and makes skip connections, which improves performance.

D. IMPROVED RECONSTRUCTION LOSS

We need the generated images which keep some crucial features from original images, such as the identity information of one person. Reconstruction loss can distinguish the difference between the authentic images and generated images through the pixels comparison. In the original StarGAN, the L1 loss function is used as a Reconstruction loss. Nevertheless, it is difficult for the L1 loss function to look for differences between tiny shifts in images and significant defects in critical areas. Besides, the calculation of the L1 loss function is relatively complicated, and there may be multiple optimal solutions. Therefore, we replace the L1 loss function with the L2 loss function. L2 loss function calculates the square of the difference between two pixels, and it will amplify two points where there is a big difference. Meanwhile, the tiny movement between the original images and generated images can only be calculated as a small difference. After squaring it, it does not expand as much as the massive difference in two corresponding pixels. In terms of gradient solution and convergence, the L2 loss function is also better than the L1 loss function. The L2 loss function is derivable everywhere, and the gradient value is also dynamically changing, which can quickly converge [50]. Therefore, in our experiment, the L2 loss function is used as reconstruction loss, instead of the L1 loss function. The formula is as follows:

$$L'_{rec} = E_{x,c,c'}[||x - G(G(x,c),c')||_2] \tag{13}$$

To improve the quality of the generated image and the similarity with original images, the Contextual Loss is added in the loss function of StarGAN. The Contextual Loss can help StarGAN retain the essential features of the original image and measure the similarity of images by comparing the pictures' high-dimensional features. The Contextual Loss is an effective and simple solution when the original image and the generated image are not entirely aligned. It compares regions with semantic similarity and considers the context of the whole image. The core idea is to treat an image as a set of features, measure the similarity between images by the similarity between features, and ignore the spatial position of features. This method makes the generated image not necessarily consistent with the original image in space. The result of feature comparison and calculation is the similarity of the two images. Contextual loss is based on context similarity to represent image similarity. Context similarity compares the cosine distance between two points in two feature graphs. The feature maps with a more similar distribution can get higher similarity. When most features in one image have similar features in another, the two images are considered to be similar. On the contrary, when most features in one image do not have similar features in the other image, the two images are considered to be not similar. The feature value of the image can be extracted by Convolution Neural Network. In this paper, VGG19 is used to extract the image features. Compared with pixel-based loss, the feature-based loss is more robust to the spatial location of the pixel. Finally, the loss function after modified is as follows:

$$L_D = -L_{adv} + \lambda_{rec} L^{r}_{cls} \tag{14}$$

$$L_G = L_{adv} + \lambda_{cls} L^{f}_{cls} + \lambda_{rec} L'_{rec} + \lambda_{CX} L_{CX}(x, y, l) \tag{15}$$

λ_{cls}, λ_{rec}, λ_{CX} is the hyperparameter of the equilibrium loss.

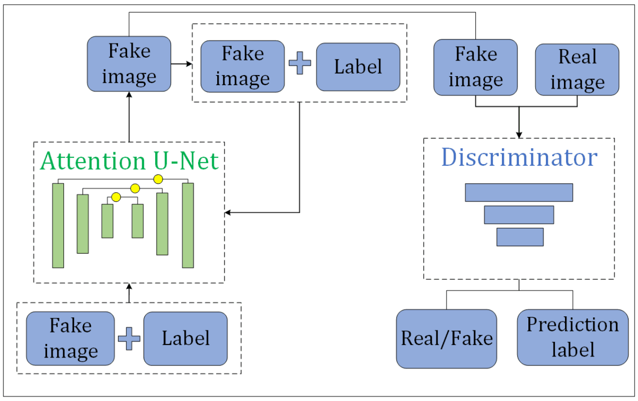


FIGURE 4. The network architecture of LAUN improved StarGAN.

E. IMPROVED GENERATOR

We use Attention U-Net as the generator in our model. The structure of LAUN improved StarGAN is shown in Fig. 4. In downsampling, it consists of repeated applications by two 3 × 3 convolutions, each followed by a ReLU and a 2 × 2 max pooling operation. The generator downsamples inputting images to get compressed features. And then, the generator upsamples the feature map to get generated images. The shallow layers can capture some simple features from pictures in the feature extraction stage. The deep layers can get some image extraction information because of the increase of the receptive field and multiple convolution operations. Upsampling has a similar structure with downsampling, and it can ensure that the final resumed feature map integrates low-level features. The features in different scales are also fused to multi-scale prediction and in-depth supervision. Four times of upsampling also makes the information of edge restore more precise. Moreover, the Attention U-Net model adds the attention mechanism, which can ignore uncorrelated areas in the image and focus on the specific important features. The feature maps from the skip connection and the previous level get into the Attention Gate to activate vital feature areas. Self-attention techniques have been put forward to eliminate the dependency on external gating information. Simultaneously, the model adopts a soft attention mechanism, which is probabilistic and utilizes standard back-propagation without the need for Monte Carlo sampling [51]. Fig. 5 shows the difference between images from StarGAN and LAUN improved StarGAN. The face on the left is generated by StarGAN, and the face on the right is generated by LAUN improved StarGAN. We can see that the quality of the mouth generated by LAUN improved StarGAN is better than that of the mouth generated by StarGAN.

FIGURE 5. The images generated by StarGAN and LAUN improved StarGAN.

IV. EXPERIMENT

In this section, we describe experiments about StarGAN and LAUN improved StarGAN. In our experiments, we used StarGAN to generated faces with different emotions. After that, we improved the StarGAN to generate higher quality images. Moreover, we used vgg16 as a classifier to get the accuracy of expression classification. The experimental learning rate is 0.0001, and the batch size is 16. The resolution of input and output images is 128 × 128 pixels, and the training step is 200000. Our experiments are trained with Adam algorithm [52]. Adam is a common optimization algorithm that can effectively reduce the training time of deep neural models. The hyperparameters about Adam in our experiments are $\beta 1 = 0.5$ and $\beta 2 = 0.999$. The device is the NVIDIA GTX2060 GPU.

A. DATABASES

The experiments are worked on the KDEF database [53] and the MMI Facial Expression database [54]. The experiment results prove the validity of our model.

1) THE KDEF DATABASE

The KDEF database has 4900 images of facial emotions, and the database contains 70 individuals. Each person displays seven different emotions with afraid, angry, disgusted, happy, neutral, sad, and surprised. Each emotion has 700 images. Also, the database contains five pan angles: −90 degrees, −45 degrees, 0 degrees, 45 degrees, 90 degrees.

2) THE MMI FACIAL EXPRESSION DATABASE

The MMI Facial Expression Database aims to provide a lot of visual data for the facial expression analysis community. The database consists of over 2900 videos and high-resolution images, which have 75 subjects. MMI Facial Expression Database has six expressions: anger, disgust, fear, happiness, sadness, and surprise. We selected 205 videos with expression tags as the training set.

B. IMPLEMENTATION DETAIL

1) EXPERIMENT ON THE KDEF DATABASE

There are seven emotions in the KDEF database, and each expression has 700 pictures. Before the experiment, a small number of images with poor quality were removed, and the final number of images was 4829. The ratio of the training set and test set is 8:2. Therefore, the training set has 3863 images, and the test set has 966 images. The pictures of the training set were put into the StarGAN, and each image generated seven different emotions. Finally, 27041 images were generated. With the original images being added, the expanded dataset has 30904 images in total.

TABLE 1. **The accuracy of different loss function on data enhancement (KDEF database).**

Method	Accuracy (%)
VGG16	93.78
VGG16 + StarGAN	94.00
VGG16 + StarGAN (L2 loss)	94.31
VGG16 + StarGAN (Contextual Loss)	94.10
VGG16 + StarGAN (L2 loss + Contextual loss)	95.03

Next, the face images and target label information of the training dataset were put into StarGAN, and the face images with target expressions could be generated. Our experiment changed the reconstruction loss function of StarGAN and added Contextual loss in the loss function. Finally, the generated images were put into VGG16 for training. We tested the model with the test dataset to obtain the facial expression recognition rates. The accuracies of the enhanced dataset are in TABLE 1. It can be seen from TABLE 1 that the recognition accuracy of VGG16 to the KDEF dataset is 93.78% before data enhancement. After using the original StarGAN for data enhancement, the accuracy of VGG16 is 94.00%, an increase of 0.22%. After adding L2 loss and contextual loss to StarGAN, the accuracy rates are 94.31% and 94.10%, respectively. After adding L2 loss and contextual loss to StarGAN, the accuracy rate is 95.03%, which is 1.03% higher than the original StarGAN. In addition, we replaced the generator in the original model with the Attention U-Net network and generated face images to enhance the data. We call it LAUN improved StarGAN. The expression recognition results are shown in TABLE 2. Finally, the accuracy rate of VGG16 after data enhancement by LAUN improved StarGAN is 95.97%, which is 0.94% higher than that with only loss function modification, 1.03% higher than original StarGAN, and 2.19% higher than that without data enhancement. To further illustrate the quality of the generated pictures, We used Fréchet Inception Distance (FID) [55] (lower is better) as the evaluation metric to measure the visual quality. FID is a common metric for evaluating images generated by GANs. It expresses the quality and diversity of generated images by comparing feature vectors between different images. The experimental results are shown in TABLE 3. It shows that LAUN improved StarGAN gets the best result. Compared with StarGAN, our method effectively improves the quality of generated images. In our experiments, it took about 15 hours to complete training the StarGAN, and about 10 hours to complete training the LAUN improved StarGAN. In the test phase, it took about 5 minutes for StarGAN to generate all images and about 5 minutes for LAUN improved StarGAN to generate all images. The result shows that our model does not increase the cost of time compared to the original StarGAN.

TABLE 2. **The accuracy of different loss function and genereator on data enhancement (KDEF database).**

Method	Accuracy (%)
VGG16	93.78
VGG16 + StarGAN (L2 loss + Contextual loss)	95.03
VGG16 + LAUN improved StarGAN	95.97

TABLE 3. **The score of FID by different models on KDEF database.**

Method	Score
StarGAN	9.11
StarGAN (L2 loss + Contextual loss)	8.63
LAUN improved StarGAN	7.62

After the original StarGAN generating images, the enhanced dataset was put into VGG16 for training. It can be known from the experimental results that the model retains the person's identity information, and can still generate different expressions in the side face state. However, some parts of generated images by the original StarGAN have defects, mainly concentrating in the mouth. In the experiment, we changed the loss function of StarGAN. The reconstruction loss was changed from L1 loss function to L2 loss function, and we added Contextual Loss in the StarGAN. Through the experiment, we could find that the problem of poor mouth generation quality has been alleviated. Meanwhile, we used the Attention U-Net instead of the original generator. The U-Net can get feature integration, and the Attention Gate can restrain the irrelevant area and highlight the important area. Therefore, the quality of image generation is further improved. After adding Contextual Loss and Attention U-Net in the model, the quality of image generation has been improved, especially for the side faces. The resulting images are shown in Fig 6. In brief, LAUN improved StarGAN

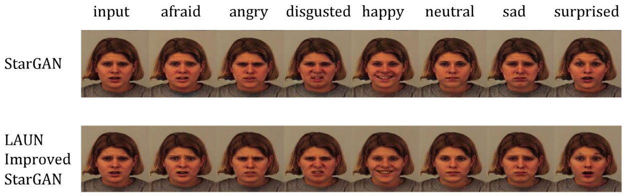


FIGURE 6. Images generated by StarGAN using different models based on the KDEF database.

TABLE 4. Comparison of experimental results with other methods on KDEF database.

Method	Accuracy (%)
VGG16	93.78
METT [56]	87.9
SCAE [57]	92.52
3-NN [58]	95
VGG16 + LAUN improved StarGAN	95.97

can produce images with better quality, and the accuracy of facial expression classification after data enhancement is also higher. TABLE 4 shows the comparison results between our model and other models.

2) EXPERIMENT ON THE MMI DATABASE

To further verify the validity of the model, we performed experiments on MMI Database. We selected 205 videos with expression labels in this dataset. Each video was selected for 20 frames, including the calm part in front of the video, the climax part in the middle of the video, and the calm part in the back of the video. A total of 4100 pictures were obtained. The division ratio is 8: 2. The training set has 3280 pictures, and the test set has 820 pictures. Each picture in the training set can generate different pictures according to different expressions. Therefore, one image can generate six pictures and a total of 19680 pictures. After putting the original images and the generated images together, the entire training set is 22960 images.

TABLE 5. The score of FID on MMI database.

Method	Score
StarGAN	26.38
StarGAN (L2 loss + Contextual loss)	23.43
LAUN improved StarGAN	18.92

We directly put the collated dataset into the StarGAN model for data generation. The generated pictures and the original training set were put together to form the training dataset. The resulting images are shown in Fig 7. VGG16 is still the classifier in this experiment. The trained classifier would be tested to get the final recognition result. After modifying the reconstruction loss function and adding Contextual loss, we experimented again to get the generated pictures and accuracy of expression classification. Finally, we modified the generator of the model to Attention U-Net to obtain the generated pictures and accuracy of expression classification. We used FID to compare the quality of the generated pictures, as shown in TABLE 5. It shows that LAUN improved StarGAN gets the best result. Same as on the KDEF database, it took about 15 hours to complete training the StarGAN, and about 10 hours to complete training the LAUN improved StarGAN. In the test phase, it took about 4 minutes for StarGAN to generate all images and about 4 minutes for LAUN improved StarGAN to generate all images. The result shows that our model does not increase the cost of time compared to the original StarGAN. Expression recognition rates are summarized in TABLE 6. The accuracy of VGG16 is 97.09% on the original MMI dataset, and 97.46% by using the original StarGAN, with an increase of 0.37%. After adding L2 loss and contextual loss in StarGAN, the accuracy rate is 97.58%, which is 0.12% higher than the original StarGAN. The accuracy of the final LAUN improved StarGAN is 98.30%, which is 0.84% higher than the original StarGAN. Compared with no enhancement, the recognition rate of the LAUN improved StarGAN is improved by 1.21%. It can be seen that after the dataset is enhanced, the expression recognition rate of the dataset has been improved. Compared with the quality of images generated by the original model, the improved model's quality of images is also better. TABLE 7 shows the comparison results between our model and other models.

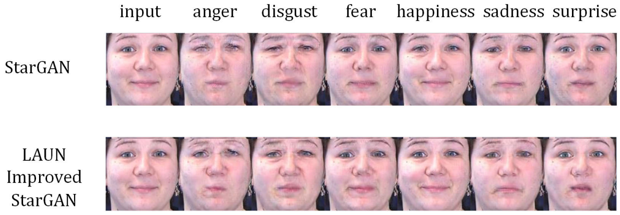


FIGURE 7. Images enhanced by different ways based on the MMI Database.

TABLE 6. The accuracy on MMI database.

Method	Accuracy (%)
VGG16	97.09
VGG16 + StarGAN	97.46
VGG16 + StarGAN (L2 loss + Contextual loss)	97.58
VGG16 + LAUN improved StarGAN	98.30

TABLE 7. Comparison of experimental results with other methods on MMI database.

Method	Accuracy (%)
VGG16	97.09
Inception-v4 [59]	94.89
Desnet201 [60]	97.20
VGG16 + LAUN improved StarGAN	98.30

V. CONCLUSION

This paper presents the LAUN improved StartGAN for data enhancement. At present, there are still insufficient and unbalanced samples in expression datasets. To alleviate this problem, we can apply the GANs network to enhance expression datasets. In addition, StarGAN can generate face images with different expressions to improve the accuracy of expression recognition. Furthermore, for solving the defects caused by StarGAN, we changed the reconstruction loss function of StarGAN and introduced the Contextual Loss. At the same time, we also turned the model generator into the Attention U-Net to improve the quality of image generation. Attention U-Net pays more attention to the important areas of pictures. The experiment results show that our model promotes the quality of generated images and classifier accuracy in emotion recognition. Meanwhile, the model can learn the identity and expression representations explicitly. Besides, the experiment also verifies that our model can generate the expression of the same person under different postures.

REFERENCES

[1] A. Mehrabian, "Communication without words," *Commun. Theory*, vol. 6, pp. 193–200, 2008.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[3] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.

[4] L. Ma, W. Chen, X. Fu, and T. Wang, "Emotional expression and micro-expression recognition in depressive patients," *Chin. Sci. Bull.*, vol. 63, no. 20, pp. 2048–2056, Jul. 2018.

[5] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 418–431, Oct. 2014.

[6] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 301–320.

[7] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "GAN-based synthetic brain MR image generation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 734–738.

[8] W. Li, M. Li, Z. Su, and Z. Zhu, "A deep-learning approach to facial expression recognition with candid images," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2015, pp. 279–282.

[9] N. Kaur and N. Bawa, "Algorithm for fuzzy based compression of gray JPEG images for big data storage," in *Proc. 2nd Int. Conf. Contemp. Comput. Informat. (ICI)*, Dec. 2016, pp. 518–523.

[10] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. Icdar*, vol. 3, 2003, pp. 1–6.

[11] S. Wang, "Facial affect detection using convolutional neural networks," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2016.

[12] M. Hu, C. Yang, Y. Zheng, X. Wang, L. He, and F. Ren, "Facial expression recognition based on fusion features of center-symmetric local signal magnitude pattern," *IEEE Access*, vol. 7, pp. 118435–118445, 2019.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

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

[15] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

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

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

[18] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–42, Apr. 2016.

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

[20] R. Hong, L. Li, J. Cai, D. Tao, M. Wang, and Q. Tian, "Coherent semantic-visual indexing for large-scale image retrieval in the cloud," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4128–4138, Sep. 2017.

[21] Q. Kai, "A painting image retrieval approach based on visual features and semantic classification," in *Proc. Int. Conf. Smart Grid Electr. Automat. (ICSGEA)*, Aug. 2019, pp. 195–198.

[22] L. Huang, C. Bai, Y. Lu, S. Chen, and Q. Tian, "Adversarial learning for content-based image retrieval," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 97–102.

[23] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Electr. Eng. III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, Nov. 1946.

[24] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, vol. 1, 1994, pp. 582–585.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

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

[27] W. Hu, J. Gao, Y. Wang, O. Wu, and S. Maybank, "Online AdaBoost-based parameterized methods for dynamic distributed network intrusion detection," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 66–82, Jan. 2014.

[28] Y. Xing and W. Luo, "Facial expression recognition using local Gabor features and AdaBoost classifiers," in *Proc. Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2016, pp. 228–232.

[29] S. Prabhakar, J. Sharma, and S. Gupta, "Facial expression recognition in video using AdaBoost and SVM," *Int. J. Comput. Appl.*, vol. 104, no. 2, pp. 1–4, Oct. 2014.

[30] F. Ren and Q. Zhang, "An emotion expression extraction method for Chinese microblog sentences," *IEEE Access*, vol. 8, pp. 69244–69255, 2020.

[31] F. Ren and Y. Zhou, "CGMVQA: A new classification and generative model for medical visual question answering," *IEEE Access*, vol. 8, pp. 50626–50636, 2020.

[32] F. Ren, W. Liu, and G. Wu, "Feature reuse residual networks for insect pest recognition," *IEEE Access*, vol. 7, pp. 122758–122768, 2019.

[33] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2015, pp. 435–442.

[34] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2121–2129.

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

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

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

[38] H. Ding, K. Sricharan, and R. Chellappa, "ExprGAN: Facial expression editing with controllable expression intensity," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–6.

[39] W. Wang, Q. Sun, Y. Fu, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y.-G. Jiang, and X. Xue, "Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 211–219.

[40] B. Bozorgtabar, D. Mahapatra, and J.-P. Thiran, "ExprADA: Adversarial domain adaptation for facial expression analysis," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107111.

[41] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

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

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

[45] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10541–10551.

[46] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 768–783.

[47] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[50] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1790–1798.

[51] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[53] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces—KDEF," CD ROM Dept. Clin. Neurosci., Psychol. Sect., Karolinska Institutet, Solna, Sweden, Tech. Rep., 1998.

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

[55] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[56] C. L. Kempnich, D. Wong, N. Georgiou-Karistianis, and J. C. Stout, "Feasibility and efficacy of brief computerized training to improve emotion recognition in premanifest and early-symptomatic Huntington's disease," *J. Int. Neuropsychol. Soc.*, vol. 23, no. 4, pp. 314–321, Apr. 2017.

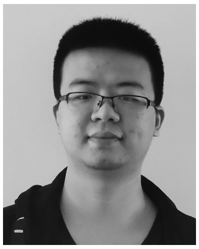
[57] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1586–1593.

[58] P. Tarnowski, M. Kolodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," in *Proc. ICCS*, 2017, pp. 1175–1184.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261

[60] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

XIAOHUA WANG received the Ph.D. degree in computer science from the Hefei Institute of Physical Science, Chinese Academy of Sciences, China, in 2005. She is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology. Her research interests include affective computing, artificial intelligence, and visual pattern recognition.



JIANQIAO GONG is currently pursuing the master's degree with the Hefei University of Technology. His research interests include image processing and facial expression recognition.

MIN HU received the M.S. degree in industrial automation from Anhui University, China, in 1994, and the Ph.D. degree in computer science from the Hefei University of Technology, Hefei, China, in 2004. She is currently a Professor with the School of Computer and Information, Hefei University of Technology. Her research interests include digital image processing, artificial intelligence, and data mining.

YU GU (Senior Member, IEEE) received the B.E. and D.E. degrees from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004 and 2010, respectively. In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar with the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor and the Dean Assistant with the School of Computer and Information, Hefei University of Technology, Hefei. His current research interests include pervasive computing and affective computing. He is a member of ACM. He was a recipient of the IEEE Scalcom2009 Excellent Paper Award and the NLP-KE2017 Best Paper Award.

FUJI REN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Sapporo, Japan, in 1991. He is currently a Professor with the Department of Information Science and Intelligent Systems, Tokushima University, Tokushima, Japan. His current research interests include natural language processing, machine translation, artificial intelligence, language understanding and communication, robust methods for dialogue understanding, and affective information processing and knowledge engineering.