Emotiongan: Facial Expression Synthesis Based on Pre-Trained Generator

Xin Ning1,2,3, Shaohui Xu2,3, a*, Yixin Zong4, Weijuan Tian2,3, Linjun Sun1,2,3 and Xiaoli Dong1,2,3,b

1 Laboratory of Artificial Neural Networks and High-speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
2 Cognitive Computing Technology Joint Laboratory, Wave Group, Beijing 100083, China
3 Beijing Wave Security Technology company limited, Beijing 102208, China
4 Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing, 100864

*a)xushaohui@wavewisdom-bj.com; b)dongxiaoli@semi.ac.cn

Abstract. Since the Generative Adversarial Networks (GANs) was proposed, researches on image generation attract many scholars’ general attention and good graces. Traditional GANs generate a sample by playing a minimax game between generator and discriminator. In this paper, we propose a new method called EmotionGAN for generating facial expression. Specifically, the inverse of the generator is firstly utilized to establish the mapping between the input and feature vector. Then the Generalized Linear Model (GLM) is used to fit the changing direction of different expressions in the feature space, which provide a linear guidance to the feature vector along the expression axis, and thus spatial distribution consistence with the target feature vector is assured. Finally the generator is applied to reconstruct the facial image of the expression. By controlling the intensity of the feature vector, the generated image can be smoothly changed on a specific expression. Experiments have shown that EmotionGAN can quickly generate face images with arbitrary expressions while ensuring identity information is not changed, and the image attributes are more accurate and the resolution is higher.

1. Introduction
In recent years, research on data generation has attracted great interest from researchers in various fields [1-2]. As one of the implementation methods for data augmentation and audiovisual entertainment, the development of GANs [3] has made great progress. Facial expression synthesis is based on the input 2D image to synthesize face images with different expressions and unchanged identity information. In this work, we use the inverse of the pre-trained generator to extract the input image latent vector, and perform expression control on the latent vector to make it have the same feature distribution as the target attributes. Finally, the target image is reconstructed by the generator.

The traditional GANs is generally considered as a minimax game between the generator and the discriminator. Regarding the synthesis of facial expression, there are roughly two ways, which are...
style-transfer and feature interpolation. Style transfer is a model for transforming images from one domain to another, such as CycleGAN[4], StarGAN[5], etc.. For style transfer, continuous transform between two discrete domains can’t be performed, and the quality of the generated images are always low. In StyleGAN [6], the method of feature interpolation is explicitly used. By linearly interpolating the latent vectors of two different images, a target image whose attributes are smoothly changed between two input images is reconstructed. However, StyleGAN can only synthesize a face image from a random vector, it can’t modify the style of the input image, and when performing feature interpolation on the two images, other unrelated attributes will also be modified. Based on the above problems, we propose a method for smoothly generating facial expressions, EmotionGAN.

EmotionGAN firstly uses a pre-trained generator to extract the latent code of the input image, afterwards, uses linear guidance to control the intensity of the latent vector, and finally restores the face image through the generator to achieve the smooth generation of the target expression. We perform experiments on 6 expression attributes, and the results show that the experimental results of EmotionGAN have achieved the state-of-the-art performance. In summary, the main contributions of this paper can be summarized as follows:

- A new method for facial expression synthesis is proposed. A pre-trained generator is used to synthesize a variety of expression images, and the task of adding attributes can be completed quickly.
- The process of expression synthesis is more controllable, and while focusing on the target area, the irrelevant area does not change.
- Using EmotionGAN for facial expression synthesis experiments, both qualitative and quantitative results show its advantages over existing models.

2. Related Work

2.1. Generative Adversarial Networks
Since Ian Goodfellow et al. proposed the Generative Adversarial Networks (GANs) [3], it has been sought after by the academic and industrial communities. Traditional GANs generally include two sub-networks: generative network and discriminatory network, and most of GANs follow this design or expand on this basis.

2.2. Image-Style-Transfer
Image style transfer is also called image translation. By transferring the style of one input image to another image, the style and content of the two images are superimposed [7]. In [4], Zhu et al. proposed CycleGAN. By using the loss of cyclic consistency to retain the information of the input image. The principle is simple, convenient and effective. However, CycleGAN can only be used for one type of transformation. When transferring between multiple domains, multiple models must be built. Based on this shortcoming, information of domain control was added to StarGAN [5]. The discriminator not only needs to learn whether the sample is authentic, but also needs to determine which domain the real image comes from. Despite this, StarGAN still can’t perform continuous transformation between two discrete domains. In GANimation [8], Pumarola et al. proposed a GANs conditionalization method based on action unit (AU) labeling, which describes the movement that defines the facial anatomy in a continuous manifold, and generating continuous expression animation from a single image and expression coding. Although the research on facial expression transfer has made great progress, there are still many problems in the accuracy and image quality of synthetic images.

2.3. Feature Interpolation
As a new method for facial style synthesis, StyleGAN [6] controls the synthesis of different attributes through linear interpolation between latent vectors, and uses a generator to restore the face image from the new latent vector. In other words, by linear interpolation between two random vectors $z_1, z_2$, 


StyleGAN can generate any style between image1 and image2. However, specifically in the experiment of facial expression synthesis, the unrelated areas between two random vectors will also be weighted, causing the style outside the target area to be changed, in addition feature interpolation can only be performed from random vectors, it’s impossible to perform facial expression synthesis on the identified image.

3. Proposed Method

In order to achieve high-quality face image synthesis with different expressions on the input image, we propose a new and highly scalable method for generating facial expressions, EmotionGAN.

![Figure 1. EmotionGAN framework.](image)

The framework of EmotionGAN proposed in this paper is shown in Figure 1. It consists of three sub-networks, namely: generative network, feature extraction network $E$ and linear regression network $L$. The weights of $G$ are shared. We use the inverse of the generator to obtain the latent vectors of the input image

$$vec = G^{-1}(X^r)$$

(1)

And then use $L$ to perform regression operations to get the direction of different expressions, and linearly guide the latent vector in different directions

$$vec^t = vec + \lambda \ast L(y)$$

(2)

Finally, synthesize various facial images with different expressions and strengths through generator $G$

$$X^t = G(vec^t)$$

(3)

3.1. Latent vector extraction

In this work, a pre-trained ResNet[9] is used to extract feature from input image, and the output feature is used to reconstruct the corresponding face image through the generator $G$. In order to ensure the similarity between the generated image and the input image, we calculated the pixel loss and perceptual loss of $X^r$ and $X^s$ respectively as a measure of the difference. By minimizing the difference, the feature extraction network is optimized to obtain the latent vector of the input image.

$$L = L_{pix} + L_{perceptual}$$

(4)

In equation (4), pixel loss guarantees consistency in image details and textures, and perceptual loss [10] makes the global structure as similar as possible. The definition of perceptual loss is as follows:

$$L_{perceptual} = \frac{1}{N} \| F(X^r) - F(X^s) \|_2^2$$

(5)

where $N$ is the size of the feature map, $N = C \ast H \ast W$, $F(\cdot)$ is the convolution operation.
3.2. Linear Regression Model

In order to explore the axis of expression generation, we use facial data with expression labels \( \{ x, y \} \), and get pairs of data between latent vectors and expression labels \( \{ \text{vec}, y \} \) through \( G^{-1} \). Since the latent space is continuous, we assume that there is an one-to-one mapping between the latent vector and the generated image. If the latent space is sufficiently decoupled, it should be possible to find a direction vector that always corresponds to each change factor \([1,1]\), and move the latent vector along the direction vector, we will get latent vectors of face images with different expressions and arbitrary strengths. Based on this, we use the Ordinary Least Squares (OLS) method to fit the direction of different expressions on \( \{ \text{vec}, y \} \), and then gradually move the latent vector of the input image along different directions. Finally, through the generator \( G \), a face image with target expression attributes is synthesized.

![Comparison of experimental results.](image)

**Figure 2.** Comparison of experimental results.

3.3. Training Details

We use tensorflow framework to experiment on ubuntu16.04 server with Xeon CPU and Nvidia 1080Ti GPU * 2, Adam [12] optimizer with \( \beta_1 = 0.9 \), \( \beta_2 = 0.999 \) and \( \epsilon = 1e-08 \) during feature extraction, the initial learning rate is 0.001, with 100 iterations, after 50 iterations, the learning rate is reduced by half.

4. Experiments

In this section, we first introduce the dataset used in the experiments, Then compare the experimental results of EmotionGAN with the results of expression transformation based on StarGAN, GANimation and Feature interpolation based on StyleGAN. Subsequently, the expression classification model was used to perform classification experiments on the synthetic images, and the analysis was performed from the qualitative and quantitative results, showing the advantages of our method.
4.1. Dataset
We perform experiments on the AffectNet [13] dataset, which contains more than one million facial images collected from the Internet, manually annotated about 440K images, and includes seven discrete expressions: natural, happy, sad, surprised, Fear, disgust, and anger. Due to the imbalanced samples of each classification, we selected a total of 40,419 images with expressions of natural, happy, sad, surprise, fear and anger. 500 images from each classification were selected as the validation set, and the remaining images were used as the training set.

4.2. Experimental Results on AffectNet

4.2.1. Qualitative results. As shown in Figure 2, from top to bottom, the 1st and 2nd rows are the results of StarGAN; the 3rd and 4th rows are the results of GANimation; the 5th and 6th rows are the results of feature interpolation; the last two rows are the results of EmotionGAN. From left to right, the 1st column are the input images with and without expression, the 2nd column is the result of neutral; the 3rd column is the result of happy; the remaining 4th, 5th, 6th and 7th columns represent angry, fear, sad and surprise results respectively.

It can be seen from Figure 2 that on the six expression attributes we selected, the quality of face images generated by the competing methods StarGAN and GANimation is not high, both are 128 * 128. The result based on StarGAN method doesn’t perform well in synthesizing face details, and it is prone to blur in the mouth area; the result based on GANimation method, expression transfer can be performed continuously between different expressions, but the quality of the output image expression depends largely on the reference image, especially for input images with expressions, the expression characteristics of output images will be inconspicuous and unrealistic will be aggravated. Emoticon images synthesized by Feature-interpolation, the irrelevant areas have also been modified a lot when changing the target attributes. Relatively speaking, the resolution of the output image of EmotionGAN is 1024 * 1024, which is more high-definition, and the output image is more in line with the real visual expression, such as happy, anger, sad, etc., and the synthetic image is more realistic.

4.2.2. Quantitative evaluation. To quantitatively evaluate the accuracy of expression synthesis by different algorithms, we first use CNN to train our own expression classification model on the training set AffectNet, and achieve a classification accuracy of 89.7% on the validation set, and use the classification model to judge the images that synthesized from StarGAN, GANimation, Feature-interpolation and EmotionGAN in testing set respectively. The results are shown in Table 1.

It can be seen from the Table 1 that EmotionGAN shows the best synthesis accuracy on the synthesis tasks of neutral, happy, angry, fear and sad facial expression. Especially in the happy expression synthesis, it can get more than 70% accuracy.

Table 1. Results of expression classification on different algorithms.

|                  | Neutral | Happy | Angry | Fear | Sad | Surprise |
|------------------|---------|-------|-------|------|-----|----------|
| StarGAN          | 48.32%  | 62.56%| 50.13%| 42.81%| 49.37%| 44.14%   |
| GANimation       | 53.54%  | 63.12%| 46.48%| 37.32%| 39.82%| 51.92%   |
| Feature          | 31.26%  | 43.96%| 37.26%| 23.71%| 37.11%| 29.84%   |
| interpolation    |         |       |       |      |     |          |
| EmotionGAN       | 62.16%  | 75.47%| 65.39%| 41.25%| 59.51%| 48.59%   |

5. Conclusion
In this article, we propose a new method for smooth generation of HD facial expression images. Compared with the existing methods, EmotionGAN can greatly improve the quality of generated images, and smoothly modify the expression of the target image without changing identity information and irrelevant areas. The innovation of this work is to use the trained generator to map the face images in the real scene to latent space, and explore the relationship between the latent vector and facial
expressions, modify the latent vector to smoothly synthesize face images with different expressions, making the synthesis process more controllable. Experimental results show that the expression synthesis based on EmotionGAN is more accurate and the resolution of the generated image is higher.

Acknowledgments
This work was supported by the National Nature Science Foundation of China, grant no. 61901436, and Shenzhen Wave Kingdom Co., Ltd.

References
[1] Dayan P, Hinton G E, Neal R M and Zemel R S 1995 The helmholtz machine Neural computation pp 889-904
[2] Gregor K, Danihelka I, Graves A, Rezende D J and Wierstra D 2015 Draw: A recurrent neural network for image generation Preprint 1502.04623
[3] Goodfellow, et al. 2014 Generative adversarial nets Advances in neural information processing systems pp 2672-80
[4] Zhu J Y, Park T, Isola P and Efros A A 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks Proceedings of the IEEE international conference on computer vision pp 2223-32
[5] Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J 2018 Stargan: Unified generative adversarial networks for multi-domain image-to-image translation Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 8789-8797
[6] Karras T, Laine S and Aila T 2019 A style-based generator architecture for generative adversarial networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4401-10
[7] Gatys L A, Ecker A S and Bethge M 2015 A neural algorithm of artistic style Preprint 1508.06576
[8] Pumarola A, Agudo A, Martinez A M, Sanfeliu A and Moreno-Noguer F 2018 Ganimation: Anatomically-aware facial animation from a single image Proceedings of the European Conference on Computer Vision pp 818-33
[9] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proceedings of the IEEE conference on computer vision and pattern recognition pp 770-8
[10] Johnson J, Alahi A and Fei-Fei L 2016 Perceptual losses for real-time style transfer and super resolution European conference on computer vision pp 694-711
[11] Ning X, Li W, Tang B and He H 2018 BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition IEEE Transactions on Image Processing pp 2575-86
[12] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization Preprint 1412.6980
[13] Mollahosseini A, Hasani B and Mahoor M H 2017 Affectnet: A database for facial expression, valence, and arousal computing in the wild IEEE Transactions on Affective Computing pp 18-31