Natural Scene Facial Expression Recognition based on Differential Features

Shenhua Hu, Yiming Hu, Jianquan Li, Yunze Chen, Mengjuan Chen and Qingyi Gu
Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

***Abstract*—As an external manifestation of human emotions, expression recognition plays an important role in human-computer interaction. Although existing expression recognition methods work perfect on constrained frontal faces, there are still many challenges in expression recognition in natural scenes due to different unrestricted conditions. Face recognition in natural scenes is a problem that the intra-class gap is larger than the inter-class gap. In order to solve this problem, we propose a method of generating a reference expression using GAN and comparing it with the original expression to generate differential features, so as to avoid interference of irrelevant information on expression recognition. Besides, we have specifically optimized the GAN network that generates reference expressions to make the generated reference expression more natural. We used Resnet50-V2 pre-trained on ImageNet to better present the differential features of the original expression and the reference expression. After testing on the two datasets, our model achieves higher accuracy than other models.**

***Index Terms*—Facial expression recognition, GAN with attention, differential feature**

I. Introduction

In recent years, with the development of image processing, computer vision, pattern recognition and artificial intelligence technology, the face-related research has been gradually subdivided into face detection, face recognition, expression recognition, and three-dimensional face reconstruction. As an external manifestation of human emotions, expression recognition has played an important role in human-computer interaction. Therefore, the in-depth study of the expression recognition task will greatly promote the development and application of the face analysis. As an important research direction in the field of bionic vision, the accuracy of expression recognition tasks under complex conditions still needs to be improved. Due to the particularity of the face, facial expression recognition still faces many challenges, such as face shape changes, illumination changes, scale changes, occlusion changes, and changes in identity information. The complex uncontrolled scenes lead to the poor stability of expression recognition, which reduces the application value of expression recognition technology. How to achieve high performance and efficient expression recognition has become an important research topic.

As early as 1976, Ekman et al. first proposed the facial behavior coding system FACS, and defined six standard expressions of the face through the action unit AU (Action Unit), including happy, angry, disgusted, sad, scared and surprised. These six expressions still have some limitations, however, due to their wide influence, most of the existing datasets are labeled with these six expressions. Although many facial expression recognition systems have been proposed and implemented, most of them are built on images which are captured in controlled environment, such as CK+, Oulu-CASIA-VIS, and other lab-collected Datasets. The expressions in these datasets are collected in a controlled environment with constant illumination, angles, and no occlusion. Besides, the expressions are more exaggerated. Common expression recognition systems often perform perfect on these datasets. However, in natural scenes, people's expressions are less noticeable, and the illumination and angles vary greatly. The performance of these expression recognition systems are not satisfactory. The researchers also make great efforts on collecting large-scale facial expression datasets in the wild, such as RAF-DB [1], AffectNet [2]. How to achieve high accuracy on these expression recognition datasets becomes a challenge.

In the field of pattern recognition, expression recognition is a typical problem that the intra-class distance is greater than the inter-class distance. The difference between different faces, angles and lighting will be much greater than the difference between different expressions of the same person. This makes it difficult for the universal classification network to learn subtle differences between different expressions — they are masked by differences between different faces. In order to solve this problem, we propose the concept of differential features, that is, we use the same neutral network to extract features of the image to be tested and the reference image, then the obtained features are subtracted, and differential features are used for expression classification. In this way, extraneous interference can be filtered out, leaving only meaningful features. In order to obtain the reference image, we are inspired by GANimation [3] to design an expression recovery network, using GAN to generate the reference image corresponding to the input image. Then use the feature extraction network to extract features of the original image and the generated image. Finally, the features are subtracted to obtain information related to the expression.

II. Related Works

We reviewed previous works considering two aspects, i.e., using image information for expression recognition and using reference images for expression recognition.

Corresponding Author: Qingyi Gu. (Email: qingyi.gu@ia.ac.cn)

A. Expression recognition with image information

The expression recognition method can be divided into geometric feature based method, apparent feature based method, and blending method according to the feature type.

Geometric feature-based method uses the shape model of the face and the face feature points to represent the expression of face, and classifies the expression by extracted geometric features of the face. Kotisa et al. proposed a face feature point localization method based on the Kanade-Lucas-Tomasi tracker, which uses the position change value of the face feature point as a feature for the classifier's expression classification. Sebe et al. [4] proposed tracking the face feature points manually labeled by Piecewise Beziervolume Deformation to extract facial expression features. Choi et al. proposed a method for expressing expression using active superficial model AAM and multi-layer perceptron. Zheng et al. [5] proposed using multi-angle facial features to express facial expressions, extracting features from each scale and angle through multi-scale segmentation of human faces, and constructing features and labels using group-based sparse descending rank regression models. The geometric feature-based method has a high dependence on the accuracy of face feature point location and tracking. Once the feature point location has large error, it will lead to expression recognition error.

The apparent feature based method uses the intensity features, texture features and gradient features of the extracted image to represent the face globally and locally. Among them, the global feature representation includes: gradient histogram HOG feature, Haar feature, LBP feature, etc. Local feature representation includes Mean intensity, Eigenimages and so on. Donato et al. proposed a facial expression recognition method based on different representation features of human face, including independent component analysis, principal component analysis, linear discriminant analysis and Gabor wavelet analysis. Zhong et al. [6] proposed a method for realizing expression recognition by using special face regions for learning different expressions. This method effectively extracts special regions of faces and extracts expression features only for special regions. The method based on apparent features has high requirements on the stability of feature, so it is difficult to ensure the accuracy of expression recognition in complex scenes.

In recent years, the method based on deep learning framework has obvious advantages in feature extraction, pattern classification, target detection and recognition. Because expression recognition can be regarded as classification problem, it is suitable to be implemented in deep learning framework. Liu et al. [7] proposed using an enhanced deep Belief Network to learn features. As the training time is accumulated, the classifier's classification ability is continuously enhanced. Khorrami et al. [8] proved that the method based on convolution neural network can improve the accuracy of expression recognition through theoretical derivation, network structure analysis and contrast experiments.

B. Expression recognition with reference image

Since the difference between different people is often larger than the difference between different expressions, it is a more effective method to perform expression recognition on an individual basis. However, there exists a problem in this method. For a image to be tested, it's hard to obtain other expressions of the person in the image. The solution to this problem is to generate other expressions of the person to be tested through generative model.

Kim et al. first introduced generative model in expression recognition. This method attempts to explain the difference between different expressions by confronting the representation. However, this method leads to the loss of facial expression features in the process of deep encoding and decoding, resulting in incomplete optimization of network parameters.

With the deepening of research on generative networks and adversarial networks, significant progress has been made in automatically animating facial expressions from a single image. DCGANs and Style GAN can generate faces hard to distinguish real from imitation. However, GAN is always lacking the actual application. In 2018, Zhang et al. [9] proposed a facial expression recognition network based on GAN. The method constructs a GAN by inputting a sample, an expression and a pose tag that generates the corresponding sample with the specified expression and pose. In this way, the dataset is expanded to make network training better. However, this model cannot ensure the identity of the generated faces unchanged. In 2018, Albert Pumarola et al. [3] proposed GANimation solving the problem of zhang. They use a GAN to generate a fake map and then use the same GAN to restore the fake image. In this way, the l_1 loss can accurately measure the accuracy of generator. At the same time, they introduced the expression loss of the fake image and the expression loss of the restored image to train the generator, making the expression of GAN more realistic. In order to generate images without changing the information irrelevant to the expression, they also introduce a attention mechanism to determine the proportion of the original image and the new image in the generated image through the generated mask. This will avoid changing some fixed things such as jewelry, glasses, etc.

III. PROPOSED METHOD

Our model is divided into three parts, face detection module, reference expression recovery module and expression recognition module, as shown in Fig. 1. The face detection module is used to detect the face in the input expression and intercept the face to avoid background interference. The reference expression generation module is used to generate a reference expression corresponding to the original image, that is, a calm expression. The expression recognition module uses a feature extraction network to extract the features of the reference expression and the expression to be tested, and then subtracts the two features to obtain a differential feature, so as to avoid interference of identity information. Finally, a multi-layer perception is used to classify the differential feature. Next we will introduce each module one by one.

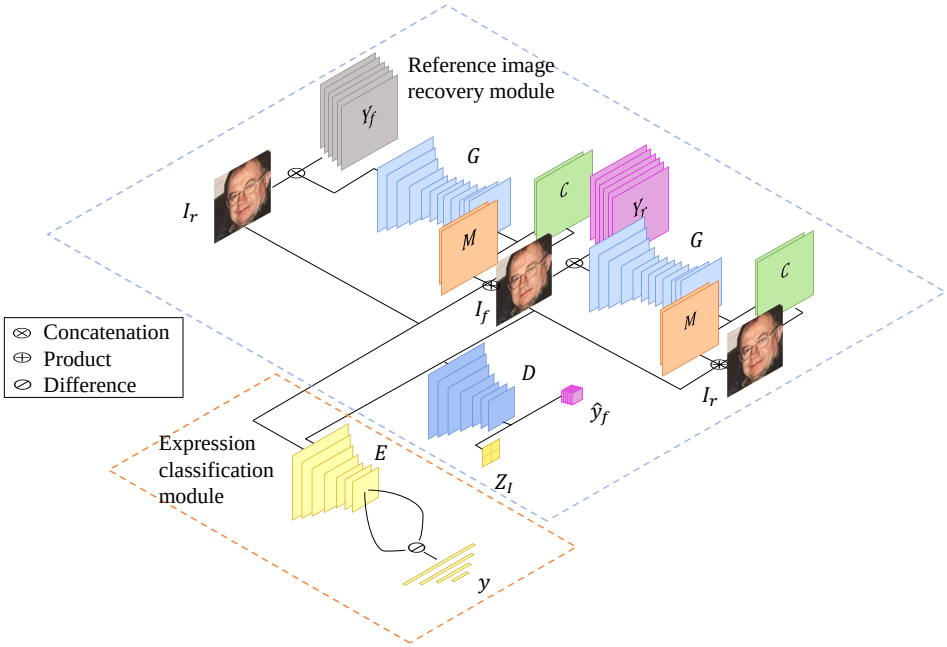

Fig. 1. Overview of our approach

A. Face detection module

We use the ResNet50-V2 pre-trained on ImageNet as the backbone of face detection module. The input to the network is original image, and the training network predicts the coordinates of the top left and bottom right corners of the picture. Then, according to the coordinates of these two points, the face is intercepted from the original image.

B. Reference image recovery module

The reference expression generation module is composed of a generator $\mathbf{G}$ and a discriminator $\mathbf{D}$. We have a face sample $\mathbf{I_r}$, the corresponding expression is a one-hot code $\mathbf{y_r}$ of length c, where c is the number of expressions, and the random expression is a one-hot code $\mathbf{y_f}$ of the same length. We stack $\mathbf{y_f}$ into $\mathbf{Y_f} \in \mathbb{R}^{w \times h \times c}$, then concatenate it with the input image to get $(\mathbf{I_r}, \mathbf{Y_f}) \in \mathbb{R}^{w \times h \times (c+3)}$ and enter them into the generator. In order to keep the generated face consistent with the original face identity, we introduced an attention mechanism when generating images. The output of generator $\mathbf{G}$ is divided into two parts, one is mask $\mathbf{M}$ and the other is color map $\mathbf{C}$. The final output can be expressed as:

$$I_f = (1 - M) \cdot C + M \cdot I_r \tag{1}$$

This can only change pixels related to expression in original image, while leaving the unrelated pixels unchanged.

The discriminator $\mathbf{D}$ has two functions, one is to judge if the image is real, the other is to classify the expression of the image. First, the discriminator draws on the idea of PatchGAN proposed by Demir et al. [10], mapping the input image $\mathbf{I}$ to a matrix $\mathbf{Z_I} \in \mathbb{N}^{w/2^6 \times h/2^6}$, and $\mathbf{Z_I}[i, j]$ represents the probability that each small block is true. Then use the idea of WGAN-GP, take the expectation of $\mathbf{Z_I}[i, j]$ as the loss function to judge the authenticity of the image. At the same time, in order to verify the expression of the image, on top of it we add an auxiliary regression head that estimates the expressions $\bar{y} = (\bar{y}_1, ..., \bar{y}_c)^T$ in the image.

C. Expression classification module

The image to be tested and the reference expression image are sent into the same feature extraction network $\mathbf{E}$ to extract image features. Since the two images are the same except for the expression, the extracted features are only different in the place related to expression. Subtracting the two features removes features that are not related to the expression, leaving only the features associated with the expression. This feature is then sent to a classification network to output the final predicted results.

IV. Training the Model

Our expression detection model is divided into three networks, so we train each of the three modules separately.

A. Face detection module

In order to get a face image for reference expression generation module, we first need to train a face detection network. The face detection network outputs the coordinates of the upper left and lower right corners of the detected face.

Expression Detection Loss. In order to speed up the training, we hope that when the predicted coordinates are far away from the real coordinates, the loss will become larger, so the loss function is a quadratic form; for detection accuracy, we hope that when the predicted coordinates are close to real coordinates, the error will not be reduced, so the loss function

is in the form of an absolute value. Therefore we use Huber loss as loss function, which is defined as:

$$L_b(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta, \\ \delta \cdot (|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{2}$$

B. Base expression generation module

The loss function we defined including four terms: image adversarial loss, which is used to push the distribution of generated images to the distribution of real images; attention loss, which is used to keep the mask smooth and avoid saturate; expression loss, which is used to force the expression of generated image similar to the desired one; and identity loss, which is used to preserve person's identity.

Image Adversarial Loss. To train the parameters of generator $\mathbf{G}$, we used WGAN-GP algorithm. Let $\mathbf{I_r}$ be the input image with the initial expression $\mathbf{y_r}$, $\mathbf{y_f}$ the desired final expression, $\mathbb{P}_\mathbf{r}$ the data distribution of the input image, and $\mathbb{P}_{\overline{I}}$ the random interpolation distribution. Then, the discriminator loss $L_D(G, D, \mathbf{I}_r, \mathbf{y}_f)$ we use is:

$$\begin{aligned} &\mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[D\left(G\left(\mathbf{I}_r|\mathbf{y}_f\right)\right)\right] - \mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[D\left(\mathbf{I}_r\right)\right] + \\ &\lambda_{\mathrm{gp}}\mathbb{E}_{\widetilde{I} \sim \mathbb{P}_{\overline{I}}}\left[\left(\left\|\nabla_{\widetilde{I}} D_{\mathbf{I}}(\widetilde{I})\right\|_2 - 1\right)^2\right] \end{aligned} \tag{3}$$

Attention Loss. When training the model, we do not have the ground-truth of mask $\mathbf{M}$ and color map $\mathbf{C}$, which are learned from the resulting gradients of the critic module and the rest of the losses. But $\mathbf{M}$ is easily saturated to 1 so that $\mathbf{I_f} = \mathbf{I_r}$, that is, the generator becomes an identity map. To solve this problem, we used regularization of l_2 weight penalty to normalize $\mathbf{M}$. At the same time, in order to make the color of the synthesized image change smoothly, we perform a *Total Variation Regularization* over $\mathbf{M}$. The attention loss $L_A(G, \mathbf{I}_r, \mathbf{y}_f)$ is:

$$\begin{aligned} &\lambda_{\mathrm{TV}}\mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[\sum_{i,j}^{h,w}\left[(\mathbf{M}_{i+1,j} - \mathbf{M}_{i,j})^2 + (\mathbf{M}_{i,j+1} - \mathbf{M}_{i,j})^2\right]\right] \\ &+ \mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[\|\mathbf{M}\|_2\right] \end{aligned} \tag{4}$$

Expression Loss. When lowering image adversarial loss, we have to ensure that our generated expressions are our target expressions. This loss consists of two parts. The loss of the expression of the real image is used to train discriminator $\mathbf{D}$, and the expression loss of fake images is used to train generator $\mathbf{G}$. This loss $L_y(G, D, \mathbf{I}_r, \mathbf{y}_r, \mathbf{y}_f)$ is:

$$\mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[\|D\left(G\left(\mathbf{I}_r|\mathbf{y}_f\right)\right)] - \mathbf{y}_f\|_2^2\right] + \mathbb{E}_{\mathbf{I}r \sim \mathbb{P}_r}\left[\|D\left(\mathbf{I}_r\right) - \mathbf{y}_r\|_2^2\right] \tag{5}$$

Identity Loss. When generating images, we also need to ensure that the identity of the generated image is unchanged. However we do not have ground-truth for supervision. Therefore, we apply the generated model to the generated image again for image recovery, and then calculate the l_1 loss between restored image and original image.

$$L_{\text{idt}}\left(G, \mathbf{I}_r, \mathbf{y}_r, \mathbf{y}_f\right) = \mathbb{E}_{\mathbf{I}_r \sim \mathbb{P}_r}\left[\|G\left(G\left(\mathbf{I}_r|\mathbf{y}_f\right)|\mathbf{y}_r\right) - \mathbf{I}_r\|_1\right] \tag{6}$$

Total Loss. To generate the target image $\mathbf{I_f}$, we build a loss function L by combine all previous losses:

$$\begin{aligned} L = &L_{\mathrm{D}}\left(G, D, \mathbf{I_r}, \mathbf{y}_f\right) + \lambda_{\mathrm{y}} L_{\mathrm{y}}\left(G, D_{\mathrm{y}}, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g\right) + \\ &\lambda_{\mathrm{A}}\left[L_{\mathrm{A}}\left(G, \mathbf{I}_{\mathbf{y}_g}, \mathbf{y}_r\right) + L_{\mathrm{A}}\left(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g\right)\right] + \\ &\lambda_{\mathrm{idt}} L_{\mathrm{idt}}\left(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g\right) \end{aligned} \tag{7}$$

whrer $\lambda_{\mathrm{A}}, \lambda_{\mathrm{y}}$ and λ_{idt} are the hyper-parameters that control the importance of every loss term.

C. Expression classification module

The original image and the reference expression image are input into the feature extraction network to extract features, and then the two features are differentiated to obtain differential feature. Finally, a classification network is applied on the obtained differential feature to obtain the expression vector y, and the cross-entropy is performed with the label t to obtain a loss function.

$$L_{cls}(t, y) = -\sum_{i=1}^{n} t_i \log\left(y_i\right) \tag{8}$$

V. EXPERIMENT

In this section, we apply our model to two publicly available natural expression datasets and compare them to the most advanced results. We selected two datasets, AffectNet and RAF-DB, which are widely adopted in the literatures.

AffectNet contains more than one million facial images collected from the Internet. Among the retrieved images, about half of them were manually annotated for the presence of seven discrete facial expressions (categorial model) and the intensity of valence and arousal (dimensional model).

Fig. 2. Original images and generated images. (Above are the original images, below are the generated images)

Real-world Affective Faces Database (RAF-DB) is a large-scale facial expression dataset with around 30,000 great-diverse facial images collected from the Internet. Each image has been labeled by about 40 independent annotators.

We show the performance of each module on two datasets. First we show the effect of the reference expression generation module, as shown in Fig. 2. As can be seen, the reference expression generation module works well. Only the parts related

to the expression are changed, and other parts (eyeglasses, background) remain as they are.

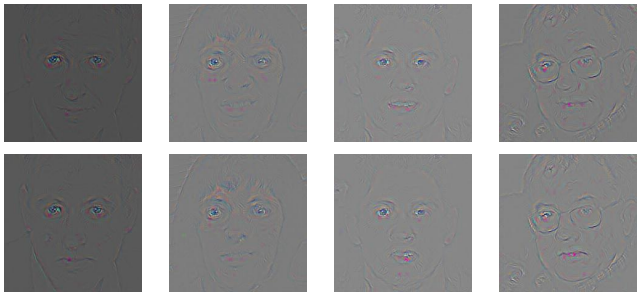

Fig. 3. Lower level features of original images and generated images. (Above are the original images, below are the generated images)

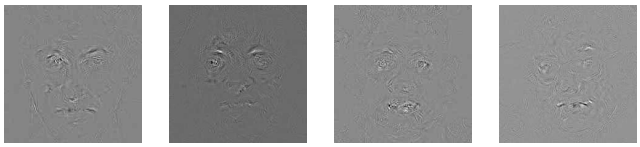

Fig. 4. Differential features.

Then we put the reference expression image and the original image into the feature extraction module. We visualized the extracted lower level features, as shown in Fig. 3. After subtracting the two features, we get the differential features Fig. 4. Then the classification network is used to classify the expression. The final classification results are shown in the following tables.

TABLE I
CLASSIFICATION RESULTS ON RAF-DB

Method	Accuracy
VGG16 [11]	80.96
DLP-CNN [12]	80.89
GAN-Inpainting [13]	81.87
pCNN [14]	81.64
gCNN [14]	83.05
Our Method	**83.55**

TABLE II
CLASSIFICATION RESULTS ON AFFECTNET

Method	Accuracy
VGG16	51.11
DLP-CNN	54.47
GAN-Inpainting	52.97
gACNN	53.90
RAN-ResNet18	53.78
Our Method	**54.40**

As we can see, our model outperforms the existing expression recognition models using a single classification network. This shows that the use of differential features can more effectively remove the interference of invalid information on expression recognition, thereby improving the accuracy of expression recognition. However, our performance on the AffectNet dataset is slightly lower than DLP-CNN. This is because the AffectNet dataset is more complex than RAF-DB dataset. It is especially important to know the structural information in advance, which leads to a slightly lower accuracy of our method.

VI. CONCLUSION

In this paper, the expression recognition belongs to the problem that the intra-class distance is greater than the inter-class distance. GAN is used to generate the reference expression of the expression to be tested, and then the features of the expression to be tested and the reference expression are compared to remove the interference of the irrelevant information on the expression recognition. However, our method still needs to be improved when the angle changes too much. Our future work is to achieve a more robust expression recovery algorithm, focusing on improving recovery performance at different angles.

REFERENCES

[1] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[3] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[4] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856–1863, 2007.

[5] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.

[6] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2562–2569.

[7] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[8] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 19–27.

[9] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3359–3368.

[10] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," *arXiv preprint arXiv:1803.07422*, 2018.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3219–3228.

[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[14] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.