Joint Pose and Expression Modeling for Facial Expression Recognition

Feifei Zhang[1,2], Tianzhu Zhang[2,3]*, Qirong Mao[1], Changsheng Xu[2,3]

[1] School of Computer Science and Communication Engineering, Jiangsu University, China

[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS [3] University of Chinese Academy of Sciences

{susanzhang, mao_qr}@ujs.edu.cn, {tzzhang, csxu}@nlpr.ia.ac.cn

Abstract

Facial expression recognition (FER) is a challenging task due to different expressions under arbitrary poses. Most conventional approaches either perform face frontalization on a non-frontal facial image or learn separate classifiers for each pose. Different from existing methods, in this paper, we propose an end-to-end deep learning model by exploiting different poses and expressions jointly for simultaneous facial image synthesis and pose-invariant facial expression recognition. The proposed model is based on generative adversarial network (GAN) and enjoys several merits. First, the encoder-decoder structure of the generator can learn a generative and discriminative identity representation for face images. Second, the identity representation is explicitly disentangled from both expression and pose variations through the expression and pose codes. Third, our model can automatically generate face images with different expressions under arbitrary poses to enlarge and enrich the training set for FER. Quantitative and qualitative evaluations on both controlled and in-the-wild datasets demonstrate that the proposed algorithm performs favorably against state-of-the-art methods.

1. Introduction

Facial expression recognition (FER) is one of the most important tasks in computer vision which plays a crucial role in numerous applications in psychology, medicine, security, digital entertainment, and driver monitoring, to name a few [41, 5, 14, 6, 3]. The main challenge of the FER is to account for large appearance changes of human faces. Despite of significant progress in recent years, it remains a difficult task for developing robust algorithms to recognize facial expression in scenarios with challenging factors such as pose variations, unconstrained facial expressions, illumination changes, and insufficient training data.

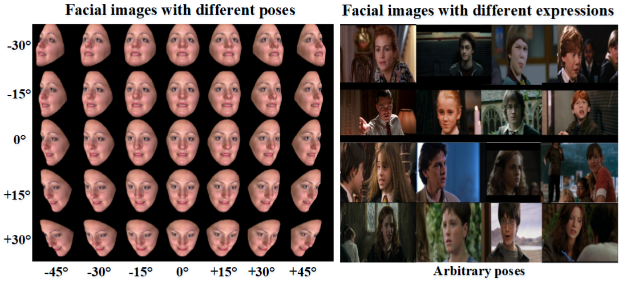

Facial images with different poses
Facial images with different expressions
-30°
-15°
0°
+15°
+30°
-45°
-30°
-15°
0°
+15°
+30°
+45°
Arbitrary poses

Figure 1. Facial expression recognition is a challenging task due to different expressions under arbitrary poses.

The facial expression recognition aims to analyze and classify a given facial image into several emotion types, such as, angry, disgust, fear, happy, sad and surprise [9]. To achieve this goal, numerous algorithms of FER [28, 21, 27] have been proposed in the literatures during the past several years. Among the existing methods, most of them are based on frontal or nearly frontal view facial images and the non-frontal or the in-the-wild facial expression recognition problem is largely unexplored. In contrast to the frontal FER, expression recognition from non-frontal facial images is challenging because it needs to deal with the issues of face occlusions, accurate non-frontal face alignment, and accurate non-frontal facial points location as shown in Figure 1. As a result, only a small part of algorithms among the proposed various methods address this challenging issue [62, 10, 58]. Different from the existing methods, we focus on the pose-invariant FER, which is to perform FER by identifying or authorizing individuals' expressions with facial images captured under arbitrary poses. Therefore, it is more challenging and more applicable in real scenarios.

However, it is not easy to perform the pose-invariant FER as shown in Figure 1. The main challenge here is to perform decoupling of the rigid facial changes due to the head-pose and non-rigid facial changes due to the expression, as they are non-linearly coupled in 2D images [66]. In details, the rigid rotation of the head results in self-occlusion, which means there is loss of information for facial expression recognition. Besides, the shape of facial texture is warped nonlinearly along with the pose change, which causes serious confusion with the inter-personal texture difference. This calls for a joint analysis of head-pose

*Corresponding Author

Table 1. The details of existing benchmarks for pose-invariant FER including the number of pose, expression, and training samples.

Dataset	Pose	Expression	Training Samples
SFEW	-	7	700
Multi-PIE	5	6	7,655
BU-3DFE	35	6	21,000

and facial expressions. Nonetheless, this remains a significant research challenge, mainly due to the large variation in appearance of facial expressions in different poses and difficulty in decoupling these two sources of variation. In order to deal with the above issues, the traditional methods usually have three distinct perspectives: (1) Extract pose-robust features as facial expression representations and employ conventional classifiers for recognition. (2) Perform pose normalization before conducting the pose-invariant FER. (3) Learn multiple classifiers for each specific poses. The success of these approaches can be attributed in good part to the quality of the feature representation used as input to the classifier. Most methods are conducted on classical hand-crafted visual features, such as local binary pattern (LBP) [64], histograms of oriented gradients (HOG) [11], and scaled-invariant feature transform (SIFT) [46], which have the limited representation power and may not handle the challenge of nonlinear facial texture warping caused by pose variation well [2, 8].

Recently, deep networks have been successfully applied on a wide range of visual tasks, such as image classification [24], object detection [11], segmentation [34], and pose estimation [33]. Inspired by the success of deep networks, an intuitive idea is to learn semantic features for the FER via a deep learning. However, deep models need to be trained with enough labeled data [23]. Thus, the first step in creating any such image classification system is gathering sufficient annotated data where each image is labeled with the correct category. For the pose-invariant FER, the publicly available datasets typically contain a very limited number of labeled samples. As shown in Table 1, there are three standard benchmarks. The Static Facial Expressions in the wild (SFEW) dataset [7] contains only 700 images (including both training and testing) while the Multi-PIE [13] has 7,655 images (5 poses and 6 expressions).

In this case, a common solution is to employ deep networks pre-trained on the ImageNet [39] and do fine tuning to further improve the feature representation power. As a result, the networks are trained separately from the FER, and the extracted features hardly benefit from the end-to-end training. End-to-end training of deep architectures is generally preferable to training individual components separately. The reason is that in this manner the free parameters in all components can co-adapt and cooperate to achieve a single objective. The other solution is to generate training data automatically. It is almost impossible to manually label training data because our goal is to perform the FER with arbitrary poses. In recent times, GAN-based approaches have been successfully used to generate impressively realistic faces [26, 20], house-numbers [60], bedrooms [35] and a variety of other image categories [15, 65] through a two-player game between a generator G and discriminator D. This inspires us to resort to the GAN to enlarge and enrich the training set. Despite many promising developments [59, 20, 29], image synthesis remains the main objective of GAN, which cannot be straightforwardly applied to facial expression recognition task.

Inspired by the above discussions, on the one hand, we design a GAN-based structure to generate facial images with different expressions and poses. On the other hand, we embed a classifier into the network to facilitate the image synthesis and conduct facial expression recognition. To disentangle the attributes (expression, pose) from the identity representation, we construct the G with an encoder-decoder structure, which serves as a facial image changer. The input to the encoder G_{enc} is a face image of any expression and pose, the output of the decoder G_{dec} is a synthetic facial image at a target expression and pose, and the learnt identity representation bridges G_{enc} and G_{dec}. Besides, we introduce two discriminators (D_{att} and D_i) into the generative adversarial network. The D_{att} is used to disentangle the pose, expression and identity from a facial image in a latent space to change the attributes (pose and expression) but retain the identity. To smooth the pose and expression transformation, the D_i is adopted to control the distribution of identity features. With an additional classifier C_{exp}, it can strive for the generated facial image to have the same expression as the input real facial image, which has two effects on G: (1) The generated facial image looks more like the input subject in terms of expression. (2) The learnt representation is more generative to synthesize an identity-preserving facial image but with different expressions and poses, and the generated facial images can facilitate the FER in turn.

The major contributions of this work can be summarized as follows. (1) We propose an end-to-end learning model by exploiting different poses and expressions jointly for simultaneous facial image synthesis and pose-invariant facial expression recognition. (2) The identity representation learning is explicitly disentangled from both expression and pose variations through the expression and pose codes in G and D. As a result, the proposed model can automatically generate facial images with an arbitrary expression under an arbitrary pose. (3) The proposed model achieves state-of-the-art facial expression recognition performance on Multi-PIE [13], BU-3DFE [51], and SFEW [7] datasets.

2. Related Work

In this section, we mainly discuss methods that are related to facial expression recognition and generative adversarial network.

Facial Expression Recognition. Extensive efforts have been devoted to recognizing facial expressions [30, 5, 54, 61, 3]. Most of existing methods on the FER study the expressions of six basic emotions including happiness, sadness, surprise, fear, anger and disgust because of their marked reference representation in our affective lives and the availability of the relevant training and test data [53]. Generally, the learning system mainly includes two stages, i.e., feature extraction and expression recognition. In the first stage, features are extracted from facial images to characterize facial appearance/geometry changes caused by activation of a target expression. According to whether the features are extracted by manually designed descriptors or by deep learning methods, they can be grouped into engineered features [10, 62, 40] and learning-based features [14, 18, 21, 27]. For the engineered features, it can be further divided into texture-based local features, geometry-based global features, and hybrid features. The texture-based features mainly include SIFT [62], HOG [11], Histograms of LBP [64], Haar features [45], and Gabor wavelet coefficients [49]. The geometry-based global features are mainly based on the landmark points around eyes, mouth, and noses [37, 38]. And the hybrid features usually refer to the features by combining two or more of the engineered features [10]. The learning-based features are based on deep neutral networks [27, 36]. Not surprisingly, almost all of them use some form of unsupervised pre-training/learning to initialize their models. It is mainly because the scarcity of labeled data prevent the authors from training a completely supervised model due to the overfitting problem. The most direct and effective solution to this problem is manually labeling more data. However, it may be infeasible for the FER with arbitrary poses. After feature extraction, in the next stage (expression classification), the extracted features are fed into a supervised classifier, e.g., Support Vector Machines (SVMs) [16], softmax [18], and logistic regression [36], to train a facial expression recognizer for a target expression. Different from existing methods, we use a variation of GAN to automatically generate facial images with different expressions and poses. Furthermore, our classifier is trained with the GAN in an end-to-end framework.

Generative Adversarial Network. In [12], Goodfellow et al. introduce the Generative Adversarial Network (GAN). They train generative models through an objective function that implements a minimax two-player game between a discriminator D - a function aiming to tell apart real from fake input data - and a generator G - a function that is optimized to generate input data (from noise) that 'fools' the discriminator. And through this game, the generator and discriminator can both improve themselves. Concretely, D and G play the game with a value function $V(D, G)$:

$$\min_G \max_D V(D,G) = E_{x\sim p_d(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

The two parts, G and D, are trained alternatively. One of the biggest issues of GAN is that the training process is unstable, and the generated images are often noisy and incomprehensible. The CGAN [31] is an extension of the GAN [12], where G and D receive an additional variable y as input. The objective function of CGAN can be rewritten as:

$$\min_G \max_D V(D,G) = E_{x,y\sim p_d(x,y)}[\log D(x,y)] + E_{z\sim p_z(z),y\sim p_y(y)}[\log(1 - D(G(z,y),y))] \tag{2}$$

This model allows the generator output to be controlled by y. Besides, during the last three years, several approaches [4, 65, 55, 25, 29] have been proposed to improve the original GAN from different perspectives. For example, the DCGAN [35] adopts deconvolutional and convolutional neural networks to implement G and D, respectively. It also provides empirical instructions on how to build a stable GAN, e.g., replacing the pooling by strides convolution and using batch normalization. More recent methods focus on incorporating constraints on the input data of generator or leveraging side information for better synthesis. For example, Mirza and Osindero [31] feed the class label to both G and D to generate images conditioned on the class label. Springenberg [44] and Luan et al. [50] generalize GAN to learn a discriminative classifier where D is trained to not only distinguish between real and fake, but also classify the images. Different from the methods [31, 44], our model can explicitly disentangle the identity representation learning from both expression and pose variations by using their codes. Compared to [50], which generates images only restricted by a discriminator, we introduce another discriminator and a content-similarity loss to make the generated facial images look like the inputs.

3. Proposed Method

In this section, we first give a brief overview of the proposed network for simultaneous facial image synthesis and pose-invariant FER. We then describe the learning process and show the difference with existing models.

3.1. Joint Pose and Expression Modeling for FER

We propose an end-to-end learning model by exploiting different poses and expressions jointly for simultaneous facial image synthesis and pose-invariant facial expression recognition. The architecture of our model is shown in Figure 2, which incorporates a generator, two discriminators, and a classifier. Before passing an image into our model, we first perform face detection using a lib face detection algorithm with 68 landmarks [52]. After the preprocessing, we feed the facial images into an encoder-decoder structured generator G to learn an identity representation. Specifically, G_{enc} learns a mapping from the input image to the identity feature representation $f(x)$. The representation is then concatenated with the expression and pose codes e and p to

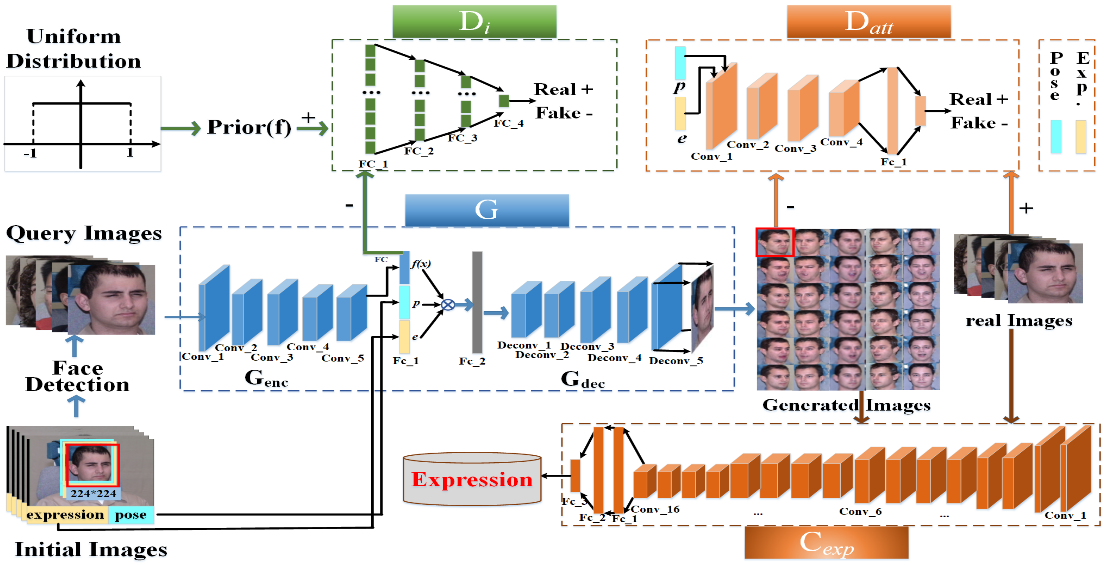


Figure 2. The overall architecture of the proposed model, which incorporates a generator G, two discriminators D_{att} and D_i, and a classifier C_{exp}. Conditioned by the expression and pose codes e and p, the proposed model can generate facial images with different expressions under arbitrary poses to enlarge and enrich the training set for the FER task.

feed to G_{dec} for face changing. Through the minimax two-player game between the generator G and the discriminator D, we can get the new labeled facial images with different poses and expressions by adding the corresponding labels to the decoder's input. Here, we use a two-discriminator structure including D_{att} and D_i. The D_{att} is to learn disentangling representations, and the other D_i is to improve the quality of the generated images. After the facial image synthesis, a classifier C_{exp} is then used to perform our FER task. We adopt a deep modeling approach for the classifier, which guarantees that, at each layer, the features become increasingly invariant to nuisance factors while maintaining discriminative information with respect to the task of facial expression recognition.

3.2. Learning

Given a facial image x with label $y = \{y^e, y^p\}$, where y^e represents the label for expression and y^p for pose, the objectives of our learning problem are threefold: (1) Synthesize a facial image $\hat{x}$ with the corresponding expression and pose labels specified by the expression and pose codes e and p. (2) Train a pose-invariant FER classifier with the generated images $\hat{x}$ and the input x. (3) Retain the identity representation with a content-similarity loss. Next we will introduce them in details.

Generator G and Discriminator D_{att}. The discriminator D_{att} is to distinguish between 'fake' images $\hat{x}$ produced by the generator G, and 'real' images from the input images x. We denote the distribution of the training data as $P_d(x)$. Conditioned by the expression and pose label y, it can help the generater G learn the disentangling representation from the facial images to change the poses and expressions but retain the identity, which is useful for our FER task, because when we generate new facial images, we just want to modify the facial expression or pose of the input x but without compromising the person's identity. The discriminator on attributes disentangling, D_{att}, and G with condition y (expression and pose) can be trained by:

$$\begin{aligned} \min_G \max_{D_{att}} & E_{x,y\sim p_d(x,y)}[\log D_{att}(x,y)]+ \\ & E_{x,y\sim p_d(x,y)}[\log(1-D_{att}(G(x,y),y)]. \end{aligned} \tag{3}$$

Generator G and Discriminator D_i. The discriminator D_i imposes the uniform distribution on the identity representation $f(x)$, which can help to smooth the pose and expression transformation. Here, the $f(x)$ is the identity representation from G_{enc}. Assuming the $Prior(f)$ is a prior distribution, and $f^* \sim Prior(f)$ denotes the random sampling process from $Prior(f)$. A min-max objective function can be used to train the G and D_i:

$$\begin{aligned} \min_G \max_{D_i} & E_{f^*\sim\, prior(f)}[\log D_i(f^*)]+ \\ & E_{x\sim p_d(x)}[\log(1-D_i(G_{enc}(x)))]. \end{aligned} \tag{4}$$

Classifier C_{exp}. The classifier C_{exp} is a task-specific loss. In the case of generation, it can be used to penalize the generator loss, which is helpful for improving the performance of the original generator G. And in the case of classification, it tries to classify the expression. We use a typical softmax cross-entropy loss for the classifier:

$$\begin{aligned} L_c(G,C) = E_{x,y^e}[& -y^e \log C(G(x),y^e) \\ & - y^e \log C(x,y^e)]. \end{aligned} \tag{5}$$

Content-similarity loss. The content-similarity loss attempts to ensure the output face sharing the expression, pose, and identity representation with the input facial image x (during training). Therefore, the input and output faces are expected to be similar as expressed in (6), where $L(:,:)$ denotes the ℓ_1 norm.

$$L_{con}(G) = L(x - G(x, y^e, y^p)). \tag{6}$$

The Objective Function. Finally, the objective function is defined as in (7) by considering the above factors.

$$\begin{aligned}\min_{G,C} \max_{D_i,D_{att}} \ & \alpha L_{con}(G) + \beta TV(G(f(x), y)) + L_c(G, C) \\ & + E_{x,y\sim p_d(x,y)}[\log D_{att}(x, y)] \\ & + E_{x,y\sim p_d(x,y)}[\log(1 - D_{att}(G(x, y))]. \\ & + E_{f^*\sim \ prior(f)}[\log D_i(f^*)] \\ & + E_{x\sim p_d(x)}[\log(1 - D_i(G_{enc}(x)))],\end{aligned} \tag{7}$$

where $TV(\textbf{.})$ denotes the total variation which is effective in removing the ghosting artifacts. The coefficients α and β balance the smoothness and high resolution. Sequentially updating the network by (3), (4), (5) and (6), we could finally learn the pose-invariant FER model.

3.3. Discussion

In this section, we show the differences of the proposed model with three most relevant GAN models including Adversarial Autoencoder (AAE) [29], disentangled representation learning-GAN (DR-GAN) [50], and conditional adversarial autoencoder (CAAE) [59]. (1) In the AAE [29], G is the encoder of an autoencoder. The AAE has two objectives in order to turn an autoencoder into a generative model: the autoencoder reconstructs the input image, and the latent vector generated by the encoder matches an arbitrary prior distribution by training D. Different from AAE, our method can explicitly disentangle the identity representation learning from both expression and pose variations by using their codes. (2) The DR-GAN [50] generalizes GAN to learn a discriminative classifier where D is trained to not only distinguish between real and fake images, but also classify real images into K classes. It is a variational autoencoder-based method mainly for disentangled representation learning for face recognition task. Different from the DR-GAN, the proposed model is mainly for generating more labeled facial images to train a deep network classifier for FER, because the training samples is the main bottleneck in facial expression recognition. Furthermore, we disentangle both the expression and pose from the facial images, and introduce a separated classifier for expression recognition. (3) The CAAE [59] extends adversarial autoencoder (AAE) to generate face images with different ages. Different from this method, our model embeds a classifier in the network and can strive for the generated facial image to have the same expression as the input real facial image.

4. Experimental Results

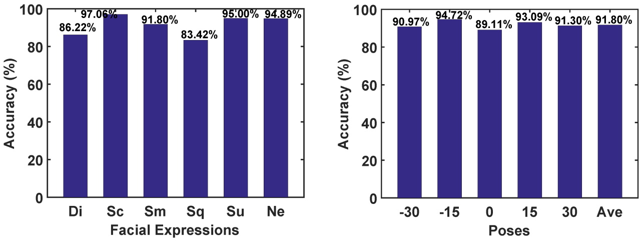
In this section, we show experimental results of our model for facial images synthesis and pose-invariant facial expression recognition. For the former task, we show qualitative results of the generated facial images under different poses and expressions. For the latter one, we quantitatively evaluate the expression recognition performance using the generated and original facial images.

4.1. Datasets

To demonstrate the effectiveness of the proposed model, we conduct extensive experiments on three standard datasets including (1) Multi-PIE [13]: the public multi-pose facial expression dataset, (2) BU-3DFE [51]: the 3D facial expression dataset, and (3) SFEW [7]: the static facial expressions in the wild dataset. The details are as follows.

Multi-PIE: The Multi-PIE is for evaluating facial expression recognition under pose and illumination variations in the controlled setting. Following the setting in [10], we use images of 270 subjects depicting acted facial expressions of Neutral (NE), Disgust (DI), Surprise (SU), Smile (SM), Scream(SC), and Squint (SQ), captured at five pan angles $-30°$, $-15°$, $0°$, $15°$ and $30°$, resulted in 1531 images per pose. Consequently, we have $1,531 \times 5 = 7,655$ facial images in total for our experiments. We perform five-fold subject independent cross-validation on the Multi-PIE. As a result, the training dataset comprises $6,124$ facial images whereas the testing one comprises $1,531$ facial images. We train the classifier using both the generated and original images, whose total number is 6124 × 5 × 6+6124=189,844.

BU-3DFE: The BU-3DFE is a 3D facial expression dataset having 100 subjects with 3D models and facial images. It contains images depicting seven facial expressions Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU) and Neutral (NE). With the exception of the neutral expression, each of the six prototypic expressions includes four levels of intensity. Following the setting in [46, 47, 48, 17], we render 2D facial images from the 3D models at the fourth level of intensity, six universal facial expressions (AN, DI, FE, HA, SA, SU), and 35 poses including 7 pan angles $(0°, \pm15°, \pm30°, \pm45°)$, and 5 tilt angles $(0°, \pm15°, \pm30°)$). Consequently, we have $100 \times 6 \times 35 \times 1 = 21,000$ face images in total for our experiments. We randomly divide the 100 subjects into a training set with 80 subjects and a testing one with 20 subjects, such that there are no overlaps between the training subjects and the testing subjects. As a result, the training set comprises $16,800$ facial images whereas the testing one



(a) Accuracy for each expression. (b) Accuracy for each pose.

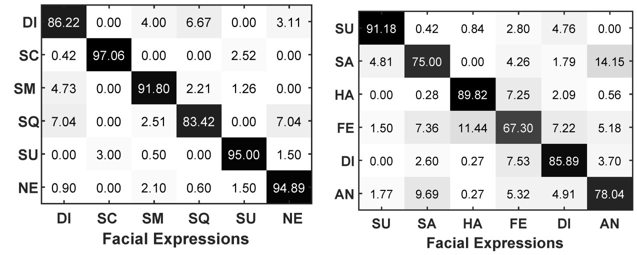
Figure 3. Overall performance on the Multi-PIE dataset.

comprises $4,200$ facial images.

SFEW: The SFEW is a dataset in the wild with 95 subjects. It consists of 700 images (346 images in Set 1, 354 images in Set 2) extracted from movies covering unconstrained facial expressions, varied head poses, changed illumination, large age range, different face resolutions, occlusions, and varied focus. The images are labeled with Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU) and Neutral (NE). We use this dataset for cross-dataset experiments. We train the model on the BU-3DFE, and test it on the SFEW. Specifically, we generate facial images with different poses and expressions on Set 1. Thus, we totally have $346 + 346 \times 7 \times 35 = 85,116$ training samples. Then we use these images to train a classifier with the same structure used on the Multi-PIE and BU-3DFE.

4.2. Implementation Details

We construct the network according to Figure 2. We first use the lib face detection algorithm with 68 landmarks [52] to crop out the faces, and resize them as 224×224. The image intensities are then linearly scaled to the range of [-1,1]. To stabilize the training process, we design the network architectures of G, D_{att}, and D_i based on the techniques in the CAAE [59]. Specifically, G is a convolutional neural network without batch normalization, and includes G_{enc} and G_{dec} that are bridged by the disentangled identity representation $f(x)$, which is the fully connected layer output in the network. Then $f(x)$ is concatenated with the expression code e and pose code p, which is a one-hot vector with the target expression y^e and pose y^p being 1. A series of fractionally-strided convolutions (FConv) [35] transforms the concatenated vector into a synthetic image $\hat{x} = G(x, y^e, y^p)$, which is the same size as the x. D_{img} and D_f is trained to optimize the object functions (3) and (4). In the discriminators D_{img} and D_f, the batch normalization is applied after each convolution layer. We adopt the VGG-Net-19 network [43] as the classifier C_{exp}. And it is trained by using the generated images $\hat{x}$ and the original images x to optimize the objective function (5). The model is implemented by using TensorFlow [1] and is trained with the ADAM optimizer [22], which is used with a learning rate of 0.0002 and momentum 0.5. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.02. The details of our architecture are included in the supplementary material.



(a) On the Multi-PIE dataset. (b) On the BU-3DFE dataset.

Figure 4. The average confusion matrix. The average recognition rate is 91.80% and 81.20%, respectively.

4.3. Quantitative Results

4.3.1 Experiments on the Multi-PIE Dataset

The overall performances over each facial expression and each pose are shown in Figure 3(a) and Figure 3(b). The average FER accuracy is 91.80% showed in the last bar in Figure3(b). A closer look at the figure reveals that, among the six expressions, there are four expressions (SC, SM, SU, and NE) with higher accuracy over 91.5%. The detailed performance of our model is provided in the confusion matrix in Figure 4(a), from which we can see that two of the most likely to be confused expressions are disgust and squint. This confusion may be due to these two expressions having similar muscle deformation around eyes.

We then evaluate our method by comparing its performance with the current state-of-the-art methods reported in [10] including kNN, LDA, LPP, D-GPLVM, GPLRF, GMLDA, GMLPP, MvDA, and DS-GPLVM. The detailed results across all views are summarized in Table 2. The mean FER accuracy is reported in the last column. The results clearly show that our method outperforms all existing methods with a 15.65% to 1.2% improvement in terms of FER accuracy. Note that all other models cannot achieve good performances in the frontal view. However, our model can significantly improve the performance attained by the generated images with arbitrary poses and expressions.

We also compare our method with the models trained by different number of generated images. Given the original N images, we can obtain $5 \times 6 \times N$ generated images. To evaluate the effect of the training data size, we randomly choose $0 \times N, 1 \times N, 5 \times N, 10 \times N, 15 \times N, 20 \times N$ images from the generated facial images during each training epoch, and then incorporate them with the original images to train the classifier, where $0 \times N$ means that the classifier is trained only using the original images. Specifically, we denote them as $0N, 1N, 5N, 10N, 15N, 20N$. The overall performance with different training samples is shown in Figure 5. It is clear that our model achieves the best recog-

Table 2. Comparison of state-of-the-art methods on the Multi-PIE dataset. The highest accuracy for each pose is highlighted in bold.

Methods	Poses					Average
	-30	-15	0	15	30	
kNN	80.88	81.74	68.36	75.03	74.78	76.15
LDA	92.52	94.37	77.21	87.07	87.47	87.72
LPP	92.42	94.56	77.33	87.06	87.68	87.81
D-GPLVM	91.65	93.51	78.70	85.96	86.04	87.17
GPLRF	91.65	93.77	77.59	85.66	86.01	86.93
GMLDA	90.47	94.18	76.60	86.64	85.72	86.72
GMLPP	91.86	94.13	78.16	87.22	87.36	87.74
MvDA	92.49	94.22	77.51	87.10	87.89	87.84
DS-GPLVM	**93.55**	**96.96**	82.42	89.97	90.11	90.60
Ours	90.97	94.72	**89.11**	**93.09**	**91.30**	**91.80**

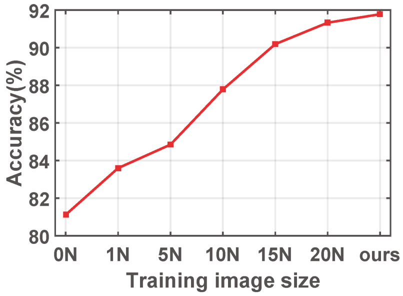


Figure 5. Effect of the number of training samples.

nition results. Besides, we can also find out that the average accuracy of the FER can be improved with the increase of the number of training samples, which further indicates the necessity of generating more labeled training samples.

4.3.2 Experiments on the BU-3DFE Dataset

The results are shown in Table 3. The rightmost column represents the average recognition error rates for different views (a total of 35 views), the bottom row represents the average recognition error rates for different facial expressions (a total of six universal facial expressions), and the bottom-right corner cell represents the average overall recognition error rate. The results show that our method achieves the average recognition accuracy of 81.20%. Furthermore, among the six expressions, surprise and happiness are easier to be recognized with accuracy over 89%. This is most likely due to the fact that the muscle deformations of both expressions are relatively large compared with others. Moreover, fear is the most difficult expression to be recognized, with the lowest at 67.30%, followed by sadness. In Figure 4(b), we show the confusion matrix for facial expression recognition by using our method. One could interpret that a contributing factor to the poor performance of fear is its confusion with happiness. This coincides with the finding of Moore and Bowden in [32], where the authors point out that the confusion is due to the expressions of fear and happiness having similar muscle deformation around the mouth. In addition, another two expressions likely to be confused are sadness and anger. These two expressions have the least amount of facial movement and thus are difficult to distinguish.

Table 3. Results on the BU-3DFE dataset in terms of the recognition rates (%). The leftmost column indicates different views (pan and tilt angles x, y in degrees), and the top row indicates different facial expressions. The highest accuracy is highlighted in bold.

Pose / Exp.	SU	SA	HA	FE	DI	AN	Ave.
$-45, -30$	90.48	66.67	90.48	61.90	85.71	71.43	77.78
$-45, -15$	100	80.95	95.24	66.67	85.71	76.19	84.13
$-45, +0$	100	80.96	90.00	76.19	76.19	76.19	83.25
$-45, +15$	90.48	80.95	90.48	85.71	90.48	76.19	85.71
$-45, +30$	89.48	70.00	87.50	80.95	90.48	71.43	81.64
$-30, -30$	100	76.19	95.24	61.90	90.00	76.19	83.25
$-30, -15$	85.71	76.19	95.24	71.43	90.48	76.19	82.54
$-30, +0$	85.71	85.71	90.48	85.71	90.48	80.95	86.51
$-30, +15$	100	80.95	90.48	76.19	90.48	71.43	84.92
$-30, +30$	85.00	66.67	85.00	76.19	90.48	61.90	77.54
$-15, -30$	90.00	76.19	90.48	66.67	80.00	80.95	80.71
$-15, -15$	85.00	80.95	95.24	61.90	80.95	76.19	80.04
$-15, +0$	80.95	80.95	95.24	71.43	80.95	80.95	81.75
$-15, +15$	90.00	76.19	90.48	71.43	90.48	85.71	84.05
$-15, +30$	90.48	71.43	85.71	76.19	90.48	80.95	82.54
$+0, -30$	89.47	76.19	71.43	57.14	80.95	57.14	72.06
$+0, -15$	90.48	75.00	90.48	57.14	85.71	85.71	80.75
$+0, +0$	90.48	76.19	90.48	66.67	76.19	85.71	80.95
$+0, +15$	100	80.00	85.71	85.71	85.71	90.00	**87.86**
$+0, +30$	89.47	68.42	82.35	70.00	90.00	85.71	80.99
$+15, -30$	90.00	71.43	85.71	66.67	85.71	80.95	80.08
$+15, -15$	90.00	71.43	90.00	57.14	76.19	85.71	78.41
$+15, +0$	90.48	76.19	95.24	66.67	76.19	85.71	81.75
$+15, +15$	100	76.19	95.00	61.90	90.48	76.19	83.29
$+15, +30$	95.00	80.95	85.71	66.67	95.00	80.95	84.05
$+30, -30$	85.00	71.43	76.19	52.38	85.00	90.48	76.75
$+30, -15$	90.48	71.43	90.00	52.38	80.95	85.71	78.49
$+30, +0$	90.48	76.19	90.48	57.14	85.71	80.95	80.16
$+30, +15$	100	80.95	95.24	61.90	85.71	85.71	84.92
$+30, +30$	95.00	80.00	95.00	66.67	85.71	76.19	83.10
$+45, -30$	90.00	66.67	90.48	52.38	90.48	57.14	74.52
$+45, -15$	90.48	71.43	95.24	52.38	90.48	80.95	80.16
$+45, +0$	85.71	76.19	95.24	66.67	85.71	61.90	78.57
$+45, +15$	90.00	61.90	90.48	76.19	90.48	76.19	80.87
$+45, +30$	82.35	65.00	83.33	71.42	80.95	80.00	77.18
Average	**91.18**	75.00	89.82	67.30	85.89	78.04	**81.20**

We then compare our method with eight previously published methods in literatures [63, 32, 62, 58, 46, 47, 48, 17]. Specifically, the methods [63, 32, 62, 58] conduct the FER on a relatively small set of discrete poses containing 5 pan angles. The algorithms [46, 47, 48, 17] use the facial images with 35 poses to train their model, which are the same as ours. Expect [58], all of other methods train their models with engineered features, such as LBP [32, 62, 17], SIFT [62, 46, 47, 48], and geometry features (83 landmark points) [63]. In [58], the SIFT feature is used as the input of DNN to learn features. Here, the model is trained separately for each step. Different from this method, ours is an end-to-end learning model. The accuracy of each model is shown in Table 4. The average FER accuracy is reported in the last column of the table. We can see that our model achieves the average recognition accuracy of 81.20%. A closer look at this table reveals that although the methods in [63, 32, 62, 58] are trained/tested on a small set of discrete poses containing only the pan rotation, our method is also competitive to the results achieved by these methods with a 1.1% to 15.2% improvement on the FER accuracy. Moreover, compared with the methods [46, 47, 48, 17], the proposed model also achieves the best accuracy (2.56% to 5.9% higher than others). This may attribute to the feature learning, which can

Table 4. Comparison of the average recognition accuracy with state-of-the-art methods for the FER on the BU-3DFE dataset.

Methods	Poses			Ave.
	tilt	pan	total	
Zheng et al. 2009 [63]	-	$(0^\circ, +90^\circ)$	5	78.3
Moore and Bowden 2011[32]	-	$(0^\circ, +90^\circ)$	5	71.1
Zheng 2014 [62]	-	$(0^\circ, +90^\circ)$	5	66.0
Zheng 2014 [62]	-	$(0^\circ, +90^\circ)$	5	78.9
Zhang et al. 2016 [58]	-	$(0^\circ, +90^\circ)$	5	80.1
Tang et al. 2010 [46]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	75.3
Tariq et al. 2013 [47]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	76.34
Tariq et al. 2014 [48]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	76.60
Jampour et al. 2015 [17]	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	78.64
Ours	$(-30^\circ, +30^\circ)$	$(-45^\circ, +45^\circ)$	35	81.20

better deal with the nonlinear facial texture warping caused by pose and individual difference.

4.3.3 Experiments on the SFEW Dataset

We finally evaluate our method on a more challenging database SFEW, in which the facial expressions are spontaneously displayed in real-world environment. As training samples in this dataset are insufficient, we adopt cross dataset experiments. Specially, we first train the generated model on the BU-3DFE dataset with 35 poses and 7 expressions (AN, DI, FE, HA, SA, SU, NE). Then we generate the corresponding facial images for the images in Set 1 in the SFEW dataset. Finally, we train the classification model on the generated and original images, and test it on Set 2.

We compare our method with five previously published methods [7, 19, 42, 10], which include the baseline obtained by the dataset creators, and four other state-of-the-art methods. The detailed results over each expression obtained from different methods are shown Table 5. The average FER accuracy is reported in the last column of the table. The difficulty of the task is further evidenced by the results in this table, where we observe a significant drop in accuracy of all methods. Overall, our method outperforms all existing methods with a 1.88% to 7.68% improvement in terms of the FER accuracy. This may attribute to the generated facial images, which can help learn discriminative features to better deal with the nonlinear facial texture warping caused by poses and individual difference.

4.4. Qualitative Results

The qualitative results of our model are illustrated in Figure 6. We randomly select a facial image from the test set, which is shown in the pink rectangle. The generated facial images with different expressions (each column) and poses (each row) are shown in the orange rectangle. And the images in the green rectangle are the ground truth. By comparing the generated images with the ground truth, it is clear that the personality has been preserved by the proposed model, and the attributes (expression and pose) have been jointly modeled in the identity representation as shown in the red rectangles. Due to limited space, more qualitative results are reported in the supplementary material.

Table 5. Comparison of the average recognition accuracy (%) with state-of-the-art methods on the SFEW dataset. The highest accuracy for each expression is highlighted in bold.

Method / Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Average
Baseline	23.00	13.00	13.90	29.00	23.00	17.00	13.50	18.90
MvDA	23.21	17.65	27.27	40.35	**27.00**	10.10	13.19	22.70
GMLDA	23.21	17.65	**29.29**	21.93	25.00	11.11	10.99	19.99
GMLPP	16.07	21.18	27.27	39.47	20.00	19.19	**16.48**	22.80
DS-GPLVM	25.89	**28.24**	17.17	42.98	14.00	**33.33**	10.99	24.70
Ours	**30.91**	21.95	19.61	**50.85**	19.23	28.00	15.52	**26.58**

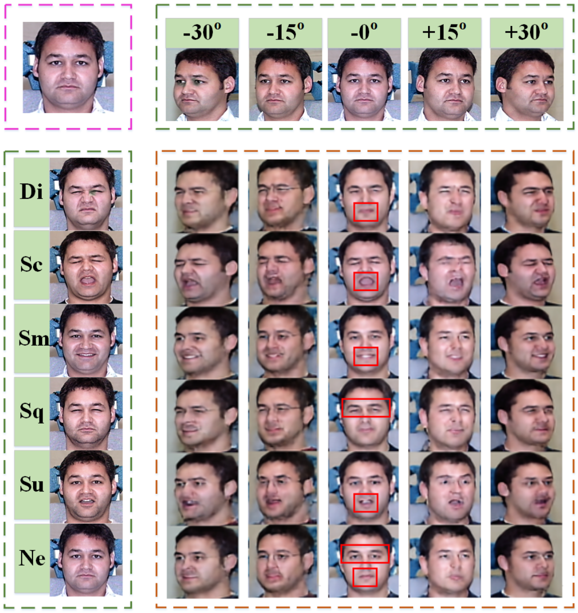


Figure 6. Example results of the generated facial images with different poses and expressions via the proposed model.

5. Conclusion

In this paper, we present an end-to-end learning model for simultaneous facial images synthesis and pose-invariant facial expression recognition. By disentangling the attributes (expression and pose) from the facial image, we can generate facial images with arbitrary expressions and poses to help train the deep neutral classification model. Experiments on three standard datasets demonstrate the effectiveness of our model. In the future, we will take other aspects in images into consideration for facial image synthesis, such as illumination, occlusion [56, 57]. The proposed model is general and can be applied to other classification tasks, such as face recognition, image classification, and audio event recognition, which we leave as future work.

6. Acknowledgment

This work is supported by National Nature Science Foundation of China under Grant 61672267, 61432019, 61532009, 61721004, 61572498, 61472379, Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, the Open Project Program of the National Laboratory of Pattern Recognition under Grant 201700022, and Beijing Natural Science Foundation (4172062).

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint:1603.04467*, 2016.

[2] Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *Computing Research Repository (CoRR)*, 1, 2012.

[3] C. F. Benitez Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016.

[4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

[5] W. S. Chu, F. D. L. Torre, and J. Cohn. Selective transfer machine for personalized facial expression analysis. *TPAMI*, 39(3):529–545, 2017.

[6] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *TPAMI*, 38(8):1548–1568, 2016.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV*, pages 2106–2112. IEEE, 2011.

[8] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology*, 7(3):37, 2016.

[9] P. Ekman and W. V. Friesen. Pictures of facial affect. In *Palo Alto,CA,USA: Consulting Psychologists Press*, 1976.

[10] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *TIP*, 24(1):189–204, 2015.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[14] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural networks. In *CVPR*, 2017.

[15] S. Gurumurthy, R. K. Sarvadevabhatla, and V. B. Radhakrishnan. Deligan: Generative adversarial networks for diverse and limited data. In *CVPR*, 2017.

[16] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel. Multi-view facial expression recognition using local appearance features. In *ICPR*, pages 3533–3536. IEEE, 2012.

[17] M. Jampour, T. Mauthner, and H. Bischof. Multi-view facial expressions recognition using local linear regression of sparse codes. In *Computer Vision Winter Workshop Paul Wohlhart*, 2015.

[18] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *ICCV*, pages 2983–2991, 2015.

[19] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *ECCV*, pages 808–821. Springer, 2012.

[20] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, pages 6089–6098, 2017.

[21] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *ICCV*, pages 19–27, 2015.

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and F.-F. Li. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, page 26, 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[25] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017.

[26] M. Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, pages 469–477, 2016.

[27] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812, 2014.

[28] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *2014 International Conference on Smart Computing*, pages 303–308. IEEE, 2014.

[29] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. In *ICLR*, 2016.

[30] Q. Mao, Q. Rao, Y. Yu, and M. Dong. Hierarchical bayesian theme models for multi-pose facial expression recognition. *TMM*, 16(4):861–873, 2017.

[31] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint: 1411.1784*, 2014.

[32] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[33] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. *arXiv preprint: 1701.01779*, 2017.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint: 1612.00593*, 2016.

[35] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[36] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, pages 808–822. Springer, 2012.

[37] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *TPAMI*, 35(6):1357–1369, 2013.

[38] O. Rudovic, I. Patras, and M. Pantic. Regression-based multi-view facial expression recognition. In *ICPR*, pages 4121–4124. IEEE, 2010.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[40] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *ACMM*, pages 357–366. ACM, 2014.

[41] E. Sariyanidi, H. Gunes, and A. Cavallaro. Learning bases of activity for facial expression recognition. *TIP*, 26(4):1965–1978, 2017.

[42] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[44] J. T. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*, 2016.

[45] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson. Generating facial expressions with deep belief nets. In *Affective Computing*. InTech, 2008.

[46] H. Tang, M. Hasegawa-Johnson, and T. Huang. Non-frontal view facial expression recognition based on ergodic hidden markov model supervectors. In *ICME*, pages 1202–1207. IEEE, 2010.

[47] U. Tariq, J. Yang, and T. S. Huang. Maximum margin gmm learning for facial expression recognition. In *FG*, pages 1–6. IEEE, 2013.

[48] U. Tariq, J. Yang, and T. S. Huang. Supervised super-vector encoding for facial expression recognition. *PR*, 46:89–95, 2014.

[49] Y. l. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *FG*, pages 229–234. IEEE, 2002.

[50] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 4, page 7, 2017.

[51] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FG*, pages 211–216. IEEE, 2006.

[52] S. Yu, J. Wu, S. Wu, and D. Xu. Lib face detection. `https://github.com/ShiqiYu/libfacedetection/`. 2016.

[53] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *TPAMI*, 31(1):39–58, 2009.

[54] K. Zhang, Y. Huang, Y. Du, and L. Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *TIP*, 26(9):4193–4203, 2017.

[55] M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *CVPR*, pages 4372–4381, 2017.

[56] T. Zhang, C. Xu, and M.-H. Yang. Learning multi-task correlation particle filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–14, 2018.

[57] T. Zhang, C. Xu, and M.-H. Yang. Robust structural sparse tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–14, 2018.

[58] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. *TMM*, 18:2528–2536, 2016.

[59] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017.

[60] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. In *ICLR*, 2017.

[61] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *CVPR*, pages 3466–3474, 2016.

[62] W. Zheng. Multi-view facial expression recognition based on group sparse reduced-rank regression. *TAC*, 5(1):71–85, 2014.

[63] W. Zheng, H. Tang, Z. Lin, and T. S. Huang. A novel approach to expression recognition from non-frontal face images. In *ICCV*, pages 1901–1908. IEEE, 2009.

[64] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, pages 2562–2569. IEEE, 2012.

[65] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.

[66] Z. Zhu and Q. Ji. Robust real-time face pose and facial expression recovery. In *CVPR*, volume 1, pages 681–688. IEEE, 2006.