Journal on Multimodal User Interfaces (2021) 15:429–440
https://doi.org/10.1007/s12193-020-00363-7

ORIGINAL PAPER

Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks

Jun He[1] · Xiaocui Yu[1] · Bo Sun[1] · Lejun Yu[1]

Received: 23 May 2020 / Accepted: 29 December 2020 / Published online: 26 January 2021

Abstract
Understanding human facial expressions is one of the key steps towards achieving human–computer interaction. Owing to the anatomic mechanism that governs facial muscular interactions, there exist powerful dependencies between expressions and action units (AUs) that are useful for exploiting such rules of knowledge to guide the model learning process. However, they have not yet been represented directly and integrated into a network. In this study, we propose a novel method for facial expressions and AUs recognition based on their dependencies on graph convolutional network. First, we train the conditional generative adversarial network to filter out identity information and extract expression information through a de-expression learning procedure. Thereafter, we apply graph convolutional network to represent dependency laying among AU nodes and embed the nodes by dividing the expression component into multi patches, corresponding to the AU-related regions. Finally, we use prior knowledge matrices to represent the dependencies between expressions and AUs and subsequently integrate them into a loss function to constrain the model. The results of our experiments indicate that such representation is effective for improving the recognition rate. They also reveal that our work achieves better performance than several popular approaches.

Keywords Facial expression · Action units (AUs) · Dependency · Conditional generative adversarial network · Graph convolutional network (GCN) · Prior knowledge

1 Introduction

Facial expression analysis refers to the differentiation of facial changes corresponding to a neutral face and is one of the most important parts in daily communication of human beings. Almost all anatomically visible facial expressions can be described by another modality known as facial action units (AUs), which refer to the local facial muscle actions, as described by the facial action coding system (FACS) [7,8]. Nowadays, owing to their applications such as in human–robot interaction, several studies [19] have focused on improving the multi-modality recognition, namely, facial expression and AU classification. These studies mainly face the following two challenges: environmental conditions, such as illumination, occlusion, low resolution, and human attribute variables, such as age, gender, and appearance. For the former, significant progress has been made [12,23,30]. However, for the latter, their changes will make human facial expressions exhibit different emotional intensities or even styles; thus, deconstructed different AU combinations. As shown in Fig. 1, a disgusted expression can be decomposed into "AU9+AU17" or "AU9+AU10+AU25". In some situations, AU17 represents the degree of expressions. Therefore, researchers propose the use of the generative adversarial network (GAN) [10] for generating the query neutral face [42] or the average face based on a database [3] to filter out identity information, then use the middle layer of the generator to classify expressions or AUs. However, based on anatomical considerations, these models ignore the symbiosis and mutual exclusion of facial AUs.

Meanwhile, because of the mapping relationship between facial expressions and AUs, some researchers started exploring the utilization of AUs for recognizing facial expressions, and vice versa. As illustrated in Fig. 1, an angry expression activates AU4, AU17, and AU23, as we generally "lower brow", "raise chin", and "tighten lid and lip" when we feel

✉ Bo Sun
tosunbo@mail.bnu.edu.cn

Xiaocui Yu
201821210003@mail.bnu.edu.cn

[1] School of Artificial intelligence, Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing Normal University, Beijing 100875, China

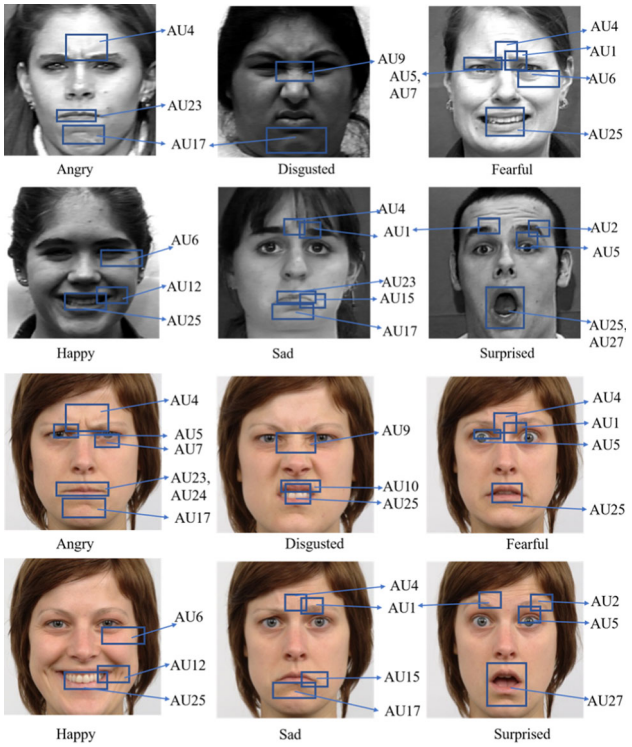


Fig. 1 Illustration of the relationship between the six basic expressions and multiple AUs from Cohn–Kanade (CK+) and Radboud Faces Database (RaFD) databases

slightly angry [6]. For the traditional methods, because the Bayesian network (BN) can imitate the process of human thinking and reasoning [2], BN is often used to model such dependencies [11]; however, its performance is limited to the input of manual features. Owing to the development of deep learning, some studies utilize the big data to obtain high-level semantic information. Through visualizing the middle layer of a convolution neural network (CNN), [13] demonstrates that the process of learning expressions is essentially learning AUs. Therefore, some models propose the construction of a multi-task [34] or multi-branch network [22] for learning each other's features. However, these methods mostly ignore the direct representation of prior knowledge as constraints for alleviating the boundary fuzzy problem of multi classification and integrate it into deep learning to simplify the network and guide the learning process.

In this paper, we introduce a novel method for the recognition of facial expressions and AUs by using the dependencies with graph convolutional network (FE-AURDGCN). In particular, we use cGAN to generate the corresponding neutral image to solve the identity-related variation problem and extract AU regions from the middle layers of the generative model. The selected AUs are regarded as nodes of the graph and we construct an AU correlation graph to represent their symbiosis relationship for learning non-geometry semantic information. Thereafter, graph convolutional network (GCN) [14] is used to guide the information propagation among nodes. Subsequently, we integrate the expressions and AUs prior distribution into the loss function to regulate network outputs and guide the training direction of the network. We conduct extensive experiments on the widely used CK+ and RaFD datasets. The results demonstrate the superiority of the proposed FE-AURDGCN framework over the state-of-the-art facial expression and AU recognition methods.

In summary, this paper has the following contributions:

(1) In this study, we formulate a novel expression and AU recognition model, known as the FE-AURDGCN model, which incorporates the cGAN and AU dependent relationship graph. This model alleviates the identity difference issue and solves the problem of encoding the appearance and geometry information of facial expressions and the relations of co-occurring of facial muscle movements.
(2) To solve the boundary fuzzy problem for some emotions and AUs classification, we propose to use their inter-dependent conditional relation matrices as prior knowledge to describe the dependencies between expressions and AUs and use them into the loss function for reducing the final error identification probability.

The remainder of this paper is organized as follows: Sect. 2 reviews related work; Sect. 3 presents the details of the FE-AURDGCN model; Sect. 4 outlines the experiments conducted in which the recognition results for AUs and expressions on the CK+ and RaFD datasets are available; finally, Sect. 5 presents the concluding remarks.

2 Related work

2.1 Expression and AU recognition

Automatic facial expression and AU recognition have garnered widespread research interest and achieved significant progress in recent years. The existing methods can be essentially divided into traditional models based on hand-crafted features and deep learning models based on neural networks. For the traditional models, apart from relying on discriminative learning methods, such as the nearest neighbor [43] and support vector machine methods [38] et al., researchers focus more on the application of prior knowledge into reasoning models. Tong et al. [41] proposed the construction of a BN to describe the dependency relationship between AUs and expressions. Li et al. [20] further developed a dynamic BN (DBN) to represent the probabilistic

relationships among facial expressions, AUs, facial components, and feature points. Wang et al. [39] used a restricted Boltzman machine to capture complex AUs relations and the correlations with expressions. Nevertheless, generalization performance of these methods is restricted because the conditional probabilities used in these topology graph models are all from target labels.

For the deep learning model, researchers have developed some effective networks by utilizing the correlation between AUs and expressions. Liu et al. [22] constructed an AU-aware depth network to enable expressions and AUs branches learn important information from each other. Zhao et al. [46] considered the anatomical attribute of facial regions and divided the face into multiple patches for AUs multi-label learning. Pons et al. [34] developed a multi-task network and the experiment revealed that simultaneous emotions classification and AUs detection can improve the expression recognition performance. Meanwhile, because of the development of the variants of CNNs known as GCN [14], Li et al. [18] used the semantic relationship between AUs as extra guidance for enhancement of facial region representation and significantly improved AU recognition performance. Liu et al. [24] used the geometric and local features of facial muscles to construct graph structure for expression classification. However, the explicit expression of the rule of knowledge based on FACS between expressions and AUs for guiding the learning direction of the network is yet to be studied.

2.2 Generative adversarial networks

Recently, the GAN has garnered increasing attention. Inspired by the adversarial idea [15], GAN [10] plays a minimax game, comprising the following two models: a generator (G) and a discriminator (D). G attempts to capture the distribution of ground truth, whereas D attempts to distinguish the generated examples from the true examples as much as possible. Owing to the development of GAN, there are increasing fields applying it or its variant, such as computer vision [36] and natural language [25]. Among them, the generation of different attributes faces is a popular topic. Liu et.al. [27] introduced a coupled GAN (CoGAN) to learn and regenerate faces with different attributes such as hair, smiling, and eyeglasses. NVIDIA [31] introduced an alternative generator architecture for generating more real faces with all types of attributes, such as freckles, pose, and even identity. Zhou et al. [37] applied conditional GAN (cGAN) to generate the neutral face from expressions and [32] used this model to recognize expressions by learning the intermediate layer of cGAN. Furthermore, Lai et al. [16] explored multi-view facial expression recognition by reconstructing the corresponding frontal face using GAN.

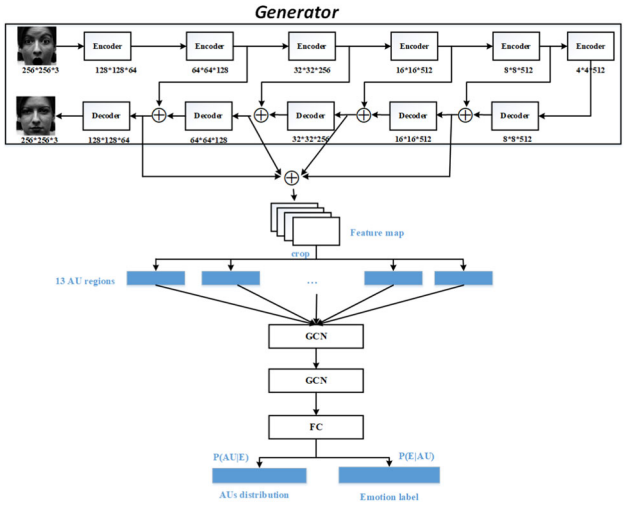


Fig. 2 Framework of our proposed FE-AURDGCN. It comprises a generative model for reconstructing a neutral face and GCN for representing the dependencies between expressions and AUs. "⊕" represents matrix concatenation along the feature dimension

3 Methodology

In this section, we introduce our FE-AURDGCN learning framework in detail. First, we briefly introduce the extraction of expression information from an identity image. Thereafter, the graph representation network of expression is presented. The overall construction of our framework is illustrated in Fig. 2.

3.1 Expression information extraction

Facial expression comprises a human face and expression information. Inspired by [42], the conditional generative adversarial network (cGAN) is used to filter out the identity information by generating the corresponding neural image of the query image. cGAN consists of G and D. In particular, the generative model generates the corresponding neutral face I_{output} through encoders and decoders and reserves the expression information in the middle layers of the network. To narrow the gap between the pseudo reconstructed face I_{output} and the ground truth I_{target} to confuse the discriminator as much as possible, we add image-difference information between them to restrain G and use L1 loss for the image similarity. The objective loss for the generator is described as follows:

$$L_{cGAN}(G) = \frac{1}{N}\sum_{i=1}^{N}\left\{-\log D\left(I^i_{input}, I^i_{output}\right) + \theta\left\|I^i_{target} - I^i_{output}\right\|_1\right\} \quad (1)$$

In other ways, the discriminator is a CNN for two classifications. Its goal is to differentiate pseudo labels $[I_{input}, I_{output}]$ from truth labels $[I_{input}, I_{target}]$. The objective loss for the discriminator is described as follows:

$$L_{cGAN}(D) = \frac{1}{N}\sum_{i=1}^{N}\left\{\log D\left(I^i_{input}, I^i_{target}\right) + \log\left(1 - D\left(I^i_{input}, I^i_{output}\right)\right)\right\} \quad (2)$$

The final loss is described in Eq. (3). The optimization terminates at a saddle point, which is a minimum for G and a maximum for D. When the model reaches equilibrium, the expression information can be extracted from the middle layers of the generative model.

$$G^* = \arg\min_G \max_D L_{cGAN}(D) + L_{cGAN}(G). \quad (3)$$

3.2 AU-related graph construction

Considering expressions of different individuals may contain different combinations of different AUs because of variation in culture and race, from the disgusted expression illustrated in Fig. 1, it is difficult to directly construct an AU relation graph to represent a specific expression. However, according to FACS, there exists a co-existent and mutually exclusive relationship between AUs caused by the mechanism of muscles. Inspired by [18], we constructed an AU-related graph to learn more semantic features through GCN [14]. GCN works by propagating information between nodes V based on the correlation matrix A. In particular, each node in the graph represents the specific AU and each value in A represents the correlation dependency of AUs. We detail the construction of A and the embedding of V as follows:

3.2.1 Correlation matrix of AUs

In this study, we define A by mining AUs co-occurrence patterns within the dataset in the form of conditional probability, i.e., $P\left(\mathrm{y}_i = 1 | y_j = 1\right)$, which denotes the probability of occurrence of AU_i when AU_j appears. Although there exist positive and negative relationships between the pairwise AUs, we only consider the influence of the positive dependencies that are interpreted in two ways as expressed in Eqs. (4) and (5). The first formula indicates that when one AU appears, the other AU is more likely to appear than not. The second formula indicates that the probability of one AU appearing when the other AU appears is higher than not. Thus, if these two conditions are satisfied, we can set $A_{\mathrm{i},j}$ as 1; otherwise, as 0 as expressed in Eq. (6). The final AUs dependent relationship is illustrated in Fig. 3.

$$P\left(\mathrm{y}_i = 1 | y_j = 1\right) > P\left(\mathrm{y}_i = 0 | y_j = 1\right) \quad (4)$$

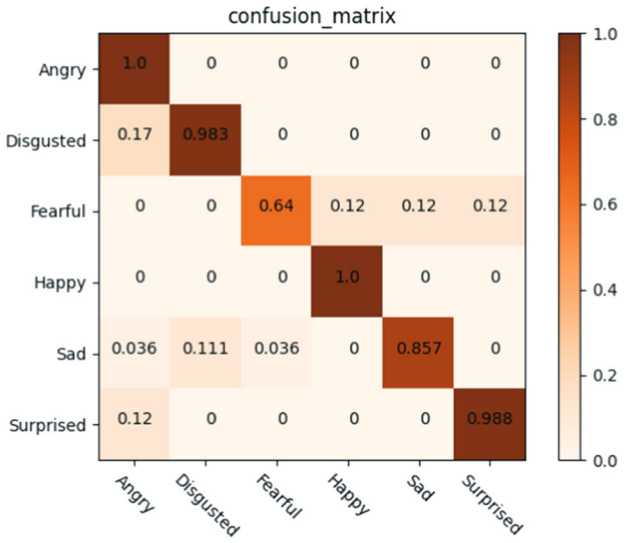


Fig. 3 AUs dependent relationship graph

$$P\left(\mathrm{y}_i = 1 | y_j = 1\right) > P\left(\mathrm{y}_i = 1 | y_j = 0\right) \quad (5)$$

$$A_{\mathrm{i},j} = \begin{cases} 1, & if\ (Eq.\ (4) = 1)\ and\ (Eq.\ (5) = 1) \\ 0, & else. \end{cases} \quad (6)$$

3.2.2 Feature embedding of nodes

According to the mapping relation between facial areas and AUs illustrated in Fig. 4, we can crop the obtained expression information into patches and set them as the corresponding node features. However, for a deep convolution network, the first one or two layers basically learn low-level features such as color, whereas the deeper layers could learn complex features such as texture [44]. Therefore, we only crop from the second layer of the generator to reduce computation cost. In other ways, we set 16*16 as the size of every AU region for the input image. Thus, based on the definition of a receptive field, the regions shrink two times after one encoder, and vice versa. Thereafter, the feature of each node is obtained by cascading each AU cropped region.

3.2.3 Convolutions on graph

To train the constructed affective graph, we perform the GCN proposed in [14]. Unlike traditional convolutions that operate on local Euclidean structures in an image, GCN uses feature descriptions X and the adjacency matrix A as inputs. The feature updating is computed as follows:

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW \quad (7)$$

$$\tilde{A} = A + I_N \quad (8)$$

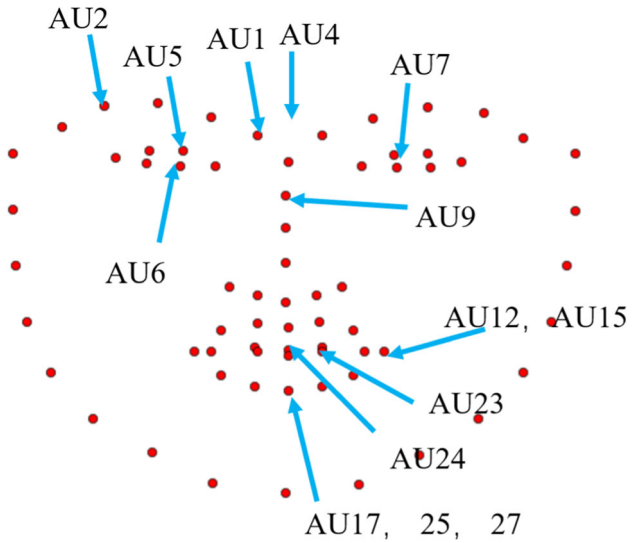


Fig. 4 Central location of facial AUs

$$\widetilde{D}_{ii} = \sum_{j} \widetilde{A}_{ij} \tag{9}$$

where Z denotes the output with $N * D^1$ dimensions, whereas $\widetilde{A}$ and $\widetilde{D}$ denote the normalized version of the correlation matrix A and D, which are computed as Eqs. (8) and (9), respectively. W denotes the learnable weight matrix with $D^0 * D^1$ dimensions. Every graph convolution layer is followed by ReLU in our experiments. Thus, we can learn and model the semantic relationships of AUs by stacking multiple GCN layers.

3.3 Loss function with prior probability

According to FACS, nearly any facial expressions can be deconstructed into the specific AUs, and vice versa. For example, when people feel angry, their face may have a higher frequency to behave as AU4, AU17, AU23, and AU24, while hardly behaving as AU1, AU25, and AU27 [5]. Thus, we can refer to such a rule as prior knowledge and use it to improve the network. It is feasible to use expression-dependent AU margin probability $P(AU|E)$ and AU-dependent expression margin probability $P(E|AU)$ to describe the constrained relation between expressions and AUs. Therefore, during the training process, we can regard the expression label and AU labels of a query image as prior knowledge and multiply them by the prior probabilities $P(AU|E)$ and $P(E|AU)$, respectively, to adjust the model outputs and then alleviate the boundary fuzzy problem.

$$L_E = -\frac{1}{N}\sum_{j=1}^{N}\left[Y\left(x_j\right)\log\left(\frac{p\left(x_j\right)\left(P_1 Q\left(x_j\right)\right)+0.05}{1.05}\right)\right] \tag{10}$$

$$L_M = -\frac{1}{NC}\sum_{j=1}^{N}\sum_{i=1}^{C}\left[Q_i\left(x_j\right)\log\left(\frac{p_i\left(x_j\right)\left(Y\left(x_j\right)P_2\right)_i+0.05}{1.05}\right) + \left(1-Q_i\left(x_j\right)\right)\log\left(\frac{1.05-p_i\left(x_j\right)\left(Y\left(x_j\right)P_2\right)_i}{1.05}\right)\right] \tag{11}$$

$$L_{\text{total}} = \lambda_1 L_E + \lambda_2 L_M \tag{12}$$

Meanwhile, the categorical and binary cross-entropy losses are often used in deep learning for the discrete multi-category and multi-label classification, respectively. To improve the training efficiency, we add prior probability into loss as demonstrated in Eqs. (10) and (11). L_E represents the expression training loss and L_M the AU training loss. Y and Q denote the ground-truth expression and AU label separately, whereas P denotes the predicted probability. Let N denote the batch size, and C denote the number of AUs. P_1 and P_2 denote the conditional probability computed according to $P(AU|E)$ and $P(E|AU)$. The total loss of the method is the combination of the losses for the emotion and AU category classification, which can be represented as Eq. (12), where the parameters λ_1 and λ_2 represent the weight coefficients for L_E and L_M, respectively.

4 Experiments

To illustrate the effectiveness of the proposed FE-AUR-DGCN, extensive experiments have been conducted on the Extended Cohn–Kanade Dataset (CK+) [26] and the Radboud Faces Database (RaFD) [17].

4.1 Experimental datasets

The CK+ dataset includes 593 sequences collected from 123 subjects. Among them, we use 309 sequences of 106 subjects that are labeled with one of six basic expressions and AUs. Each video starts with a neutral face and reaches the peak in the last frame. Hence, the apex images are selected to construct datasets and the following 13 AUs, whose frequency of occurrence are higher than 10, are used in the experiment: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU17, AU23, AU24, AU25, and AU27. Tables 1 and 2 summarize the statistical results of the conditional probabilities $P(AU|E)$ and $P(E|AU)$, which are used in the loss function. The RaFD dataset includes 8,040 images from 67 subjects. This dataset contains eight emotion expressions with three gaze directions taken from five view angles. Similar to CK+ datasets, we select images annotated with six basic expressions. Although the dataset does not provide AU labels, each model was trained by a FACS coder to exhibit each emotion. Therefore, we can set AU labels of each image as illustrated in reference [17]. In addition to the 13 AUs selected in CK+, AU10 and AU15 are both selected. Fig. 1 summarizes the

Table 1 Expression-dependent AU margin probability $P(AU|E)$

	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU12	AU17	AU25	AU24	AU25	AU27
Anger	0	0	**0.89**	0.13	0.18	0.71	0.07	0.02	**0.87**	**0.8**	**0.73**	0	0
Disgust	0	0	**0.61**	0	0.35	**0.56**	**0.98**	0.03	**0.68**	0.04	0.12	0.15	0
Fear	**0.88**	0.4	**0.84**	**0.64**	0.12	0.24	0	0.08	0.12	0	0	**0.92**	0
Happy	0	0	0	0	**0.96**	0.1	0	**0.97**	0	0	0	**0.97**	0
Sad	**0.93**	0.25	**0.82**	0	0	0.04	0	0	**0.96**	0.11	0.04	0	0
Surprise	**0.98**	**0.98**	0.01	**0.84**	0	0	0	0.04	0	0.01	0	**0.99**	**0.87**

The bold data indicates which AUs are important for expression recognition

Table 2 AU-dependent expression margin probability $P(E|AU)$

	Anger	Disgust	Fear	Happy	Sad	Surprise
AU1	0	0	0.17	0	0.2	**0.63**
AU2	0	0	0.1	0	0.07	**0.83**
AU4	**0.33**	0.3	0.17	0	0.19	0.01
AU5	0.07	0	0.17	0	0	**0.76**
AU6	0.08	0.19	0.03	**0.7**	0	0
AU7	**0.41**	**0.41**	0.08	0.09	0.01	0
AU9	0.05	**0.95**	0	0	0	0
AU12	0.01	0.03	0.03	**0.89**	0	0.04
AU17	**0.36**	**0.36**	0.03	0	0.25	0
AU23	**0.86**	0.05	0	0	0.07	0.02
AU24	**0.81**	0.17	0	0	0.02	0
AU25	0	0.05	0.13	**0.37**	0	0.45
AU27	0	0	0	0	0	**1**

The bold data indicates which expression is important for AUs recognition

specific prior distributions between facial expressions and AUs, which are used in the loss function.

4.2 Implementation details

First, for the preprocessing of the input image, MTCNN [45] and OpencCV toolbox are employed to detect a human face and extract face landmarks separately. Thereafter, we use MMI database [28] to pre-train the cGAN and fine-tune the generative model with CK+ and RaFD datasets. During the training of GCN, we set $\theta = 0.05$ and $\lambda_1 = \lambda_2 = 1$. We use an Adam optimizer with a learning rate of 0.0002 and the mini-batch size is set as 16. All models are trained using NVIDIA GeForce GTX 1080 GPU based on the tensorflow [29].

4.3 Evaluation criteria

We evaluate our method in recognition of expressions and AUs with the accuracy and F1-score, respectively. F1-score is extensively used in binary classification. It considers both the precision P and recall R, and its specific computation is described asEqs. (13)–(15).

$$F1 - score = \frac{2PR}{P + R} \tag{13}$$

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

Where TP denotes the number of true positives and FP the number of false positives. FN denotes the number of false negatives.

Finally, we set the average value of accuracy and F1-score as the overall evaluation criteria for the model performance, as expressed in Eq. (16).

$$Avg = \frac{Accuracy + F1 - score}{2} \tag{16}$$

4.4 Ablation study

*Effectiveness of cGAN*To verify the effectiveness of cGAN, we compare the performance of our proposed model to those that do not employ cGAN. In particular, we extract AU information from images directly and apply GCN to recognize AUs and expressions. As summarized in Tables 3 and 4, the effectiveness of cGAN is clear. In the CK+ dataset, our model achieves a performance of nearly 20%, 0.12 with respect to (w.r.t.) expression recognition accuracy and AU recognition F1-score, respectively, compared with the model constructed with only GCN. In the RaFD dataset, the performance is more significant, which is approximately 40%, 0.30 in terms of accuracy and F1-score. Thus, we can observe that the identity variable has a significant impact on expression and AU recognition and the employment of cGAN is effective for the tasks. Meanwhile, Fig. 5 illustrates some samples of the reconstructed neutral face using cGAN on CK+ and RaFD databases, respectively. The first column represents the input image, second column the generated face, and third column the ground truth face. The filtered expressions from the top to bottom images are as follows: angry, disgusted, fearful, happy, sad, and surprised expression images. As shown, the expression information is removed successfully by the generator while the identity information is reserved.

Table 3 Ablation study on the CK+ database

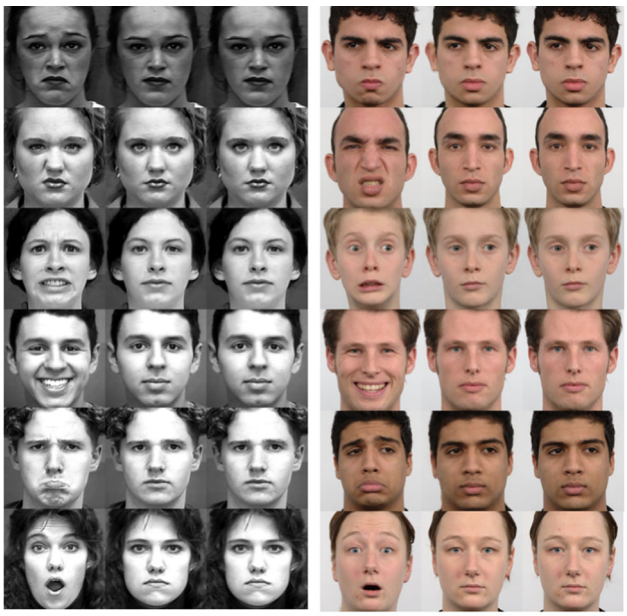
Conditions	Accuracy (%)	F1-score	Avg
Img_GCN w P_1* & P_2*	70.50	0.770	0.7375
cGAN_MLP w/o P_1 or P_2	90.30	0.890	0.8965
cGAN_MLP w P_1 & P_2	94.50	0.891	0.9180
FE-AURDGCN w/o P_1 or P_2	91.90	0.887	0.9030
FE-AURDGCN w P_1	93.50	0.891	0.9130
FE-AURDGCN w P_2	93.20	0.884	0.9080
FE-AURDGCN w P_1 & P_2	**95.10**	**0.894**	**0.9225**

The bold data indicates the best performance under different conditions
*P_1 represents $P(AU|E)$ and P_2 represents $P(E|AU)$

Table 4 Ablation study on the RaFD database

Conditions	Accuracy (%)	F1-score	Avg
Img_GCN w P_1 & P_2	50.17	0.662	0.5818
cGAN_MLP w/o P_1 or P_2	**94.61**	0.889	0.9175
cGAN_MLP w P_1 & P_2	92.37	0.954	0.9388
FE-AURDGCN w/o P_1 or P_2	92.12	0.922	0.9216
FE-AURDGCN w P_1	93.03	0.958	0.9442
FE-AURDGCN w P_2	92.70	0.927	0.9270
FE-AURDGCN w P_1 & P_2	94.28	**0.963**	**0.9529**

The bold data indicates the best performance under different conditions



CK+ RaFD

Fig. 5 Illustration of the generated face by the generator from CK+ and RaFD databses

*Effectiveness of Graph Convolutional Network*To verify the effectiveness of the GCN for emotion recognition, we compare our model to those only using multi-layer perceptron (MLP) after cGAN (cGAN_MLP). In the CK+ dataset, as summarized in Table 3, regardless of whether the proposed loss function is used, our model performs better in expressions recognition. In detail, our method achieves a performance of 1.6% compared with the cGAN_MLP model, whereas with the help of the loss function with prior knowledge, the performance of our model is higher than 0.6%. In the RaFD dataset, as summarized in Table 4, from the model without the help of P_1 and P_2, although our model performance in expression recognition is worse than that of the model with MLP, our model's performance in AU recognition is better, which is higher than 0.033. Compared with the model with P_1 and P_2, our model achieves 1.91%, 0.009 boost in both expression and AU recognition. Generally, the proposed model FE-AURDGCN performs better than the model cGAN_MLP. We can conclude that the construction of AUs-related knowledge graph is useful for expression or AU recognition.

*Effectiveness of Loss Function with Prior Knowledge*To verify the effectiveness of rule-based prior knowledge expressions, we have compared the performance of our proposed FE-AURDGCN to those without relation expressions. In the CK+ dataset, as summarized in Table 3, we can clearly observe that the model with two conditional probability constraints achieves 3.2% and 0.007 performance boost w.r.t. expression recognition accuracy and AU recognition F1-score when compared with the model without prior knowledge. In the RaFD dataset, the proposed method achieves 2.16% and 0.041 performance boost. Meanwhile, during the training process, when the prediction of some AU occur-

Table 5 Comparison of quantitative AU recognition results on CK+ database

AUs	F1-score			AUC		
	SFL-SRM [47]	LRBN [9]	Ours	WLS-RF [4]	DAUGN [24]	Ours
AU1	0.687	0.933	**0.965**	0.984	0.983	**0.987**
AU2	0.788	**0.932**	0.907	0.982	0.975	**0.992**
AU4	0.632	0.817	**0.874**	0.954	**0.956**	0.935
AU5	0.721	0.843	**0.865**	0.957	**0.962**	0.957
AU6	0.711	**0.84**	0.814	0.955	**0.967**	0.946
AU7	0.555	0.627	**0.793**	**0.902**	**0.918**	0.917
AU9	0.873	0.932	**0.95**	0.903	**0.995**	0.993
AU12	0.848	0.895	**0.939**	0.96	**0.972**	0.963
AU17	0.848	0.861	**0.879**	0.951	**0.955**	0.941
AU23	0.429	0.750	**0.824**	–	–	**0.953**
AU24	0.325	0.526	**0.800**	–	–	**0.969**
AU25	0.916	**0.97**	0.952	**0.991**	0.987	0.979
AU27	0.900	0.903	**0.921**	–	0.906	**0.976**
Avg	0.710	0.833	**0.894**	0.951	0.961	**0.962**
Avg. of Com.	**0.894**	**0.894**	**0.894**	**0.961**	**0.962**	**0.962**

The bold data indicates the best recognition performance for AUs compared with other state-of-the-art methods

rence is low while there is actually a high probability from $P(AU|E)$, the output probability would be increased by multiplying this probability; otherwise, the reverse would occur. Similarly, the final output of expressions would also be redressed. Thus, the importance of the prior knowledge expression can be observed.

4.5 Evaluation of AU recognition

For the recognition of AUs, we compare our method to alternative methods, including shared feature learning and semantic relation model (SFL-SRM) [47], latent regression Bayesian network (LRBN) [9], the confidence-weighted local subspace Random Forest (WLS-RF) [4], and deep AUs graph network (DAUGN) [24]. As summarized in Table 5, we can clearly observe that our model outperforms all of these state-of-the-art methods. SFLSRM adopts a multi-task feature learning method for learning the shared features and thereafter uses a BN to model the co-existent and mutual-exclusive semantic relations among AUs from the target labels. [9] proposes the construction of a three-layer hybrid BN, whose top two layers consist of a latent regression BN for representing relations among multiple AUs, and the bottom two layers are BNs that use expressions to facilitate the estimation of label dependencies among AUs. WLS-RF algorithm extracts a local expression subspace to describe facial expressions as well as AUs. DAUGN proposes a novel method to local AUs region and uses a graph-based CNN to combine the local-appearance and global-geometry information to recognize expressions or AUs. Compared to the first two methods, our method achieved 0.184 and 0.061 higher F1-scores, along with 0.010 and 0.001 higher AUC compared to the last two models, respectively. Although the results of our work are very close to the baseline in terms of AUC, our model can provide more information, such as more AU labels and expression labels, implying that we can obtain the same information with fewer computations and reduced time cost. On the other hand, for the specific AUs, the F1 scores of our method are higher than those of others in 10 out of 13 AUs, and the AUC are 5 out of 13 AUs. Therefore, the overall effects of our algorithms are better and demonstrate the superiority of our method over other methods.

Table 6 Comparison of quantitative expression recognition results on CK+ database

Model		Accuracy (%)
Single task	AUDN [22]	92.05
	WLS-RF [4]	94.30
	DeRL [42]	97.00
	DTAGN [12]	**97.41**
Multi task	DBN [20]	87.40
	BN [11]	91.40
	3DCNN+DAP [30]	92.40
	FE-AURDGCN (ours)	**95.10**

The bold data indicates the best performance of expression recognition or the final result of our method

4.6 Evaluation of emotion recognition

For the CK+ dataset, we compare our method to other state-of-the-art methods, including AU-aware deep networks [22], WLS-RF [4], de-expression residual learning (DeRL) [42], deep temporal geometry network (DTAGN) [12], dynamic BN (DBN) [20], BN [41], and 3D CNNs with deformable

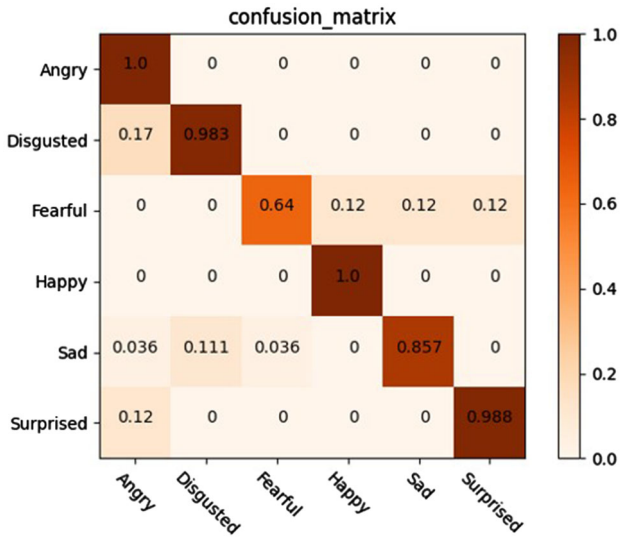


Fig. 6 Confusion matrix on CK+ database

action parts (3DCNN+DAP) [30]. As summarized in Table 6, in terms of multi-task recognition results, the average accuracies of our model FE-AURDGCN showed improvements of 7.70%, 3.70%, and 2.70%. DBN constructs a three-layer model with facial expressions, AUs, facial features, and landmark points, whereas BN adds two prior layers known as brain cognition and facial muscles layers on this basis. However, the inputs of BN are manual characteristics, which cannot be learned from end-to-end. 3DCNN+DAP uses 3D filters from local action parts to predict the expression intensity for a video segment. To evaluate the practicability, we also compare our model with single-task methods. The results indicate that our multi-task model shows improvements of 3.05% and 0.80% over AUDN and WLS-RF, respectively, while its performance is 1.9% and 2.3% lower than DTAG and DeRL, respectively. AUDN generates a complete representation of facial images to expressly describe the appearance in a specific area; however, it only considers partial patches rather than the entire face. DTAGN uses the temporal information extracted from videos and utilizes other models to fine-tune network parameters, while our model only employs static images to recognize expressions. Our model is more suitable for several applications where sequences are not available. DeRL requires a significant amount of data to train because its training result has a significant and direct influence on the results, while our model pre-trains cGAN based on only a small dataset without data augmentation, which requires fewer computations. Moreover, our model is a multi-task network, which means that we can provide more detailed expression information, such as AUs using fewer newwork parameters and at a reduced time cost. On the other hand, the application of the prior knowledge, which includes the AU dependency relationship and the mapping relationship between expressions and AUs,

Table 7 Comparison of quantitative expression recognition results on RaFD database

Model		Accuracy (%)
Single task	NNE [1]	93.75
	SURFB [35]	92.00
	SVM [21]	94.51
	MCCNN [33]	**98.17**
	TLCNN [40]	97.75
Multi task	**FE-AURDGCN (ours)**	**94.28**

The bold data indicates the best performance of expression recognition or the final result of our method

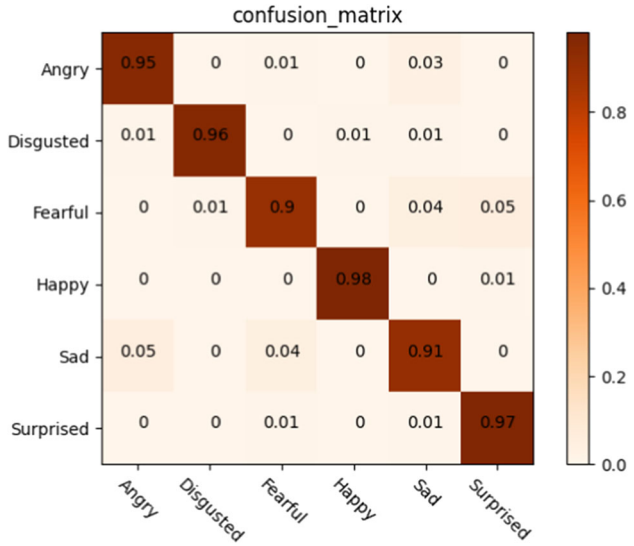


Fig. 7 Confusion matrix on RaFD database

accords with the human psychological mechanism. In this sense, our model has considerable value in further exploring research. Meanwhile, the confusion matrix in Fig. 6 demonstrates that with the help of a graph structure, the highest accuracy reaches up to 100% at angry and happy expressions. However, because of excessively small sample sizes, there exist significant errors in fearful and sad expressions.

For the RaFD dataset, as summarized in Table 7, we compare our method to other state-of-the-art methods, including the neural network ensemble (NNE) [1], SURF boosting (SURFB) [35], SVM [21], multi-channel CNN (MCCNN)) [33], and transfer learning convolution network (TLCNN) [40]. Because of the lack of AU labels, there only exist a single-task model to recognize emotions. Although MCCNN and TLCNN perform better than our model, our model is more practical. MCCNN learns and fuses the spatial-temporal features known as optical flow, but the query neutral faces are not always available. TLCNN requires a large dataset to pre-train the deep network and subsequently fine-tune it to achieve expression recognition; in contrast, our network has fewer layers. Moreover, our model provides additional AUs information, which offers some

Table 8 Recognition results of the proposed method in cross-database experiments

Database	Model	Accuracy (%)	F1-score
From RaFD to CK+	P_1	42.07	0.434
	P_2	45.31	0.346
	P_1 & P_2	**54.05**	**0.624**
From CK+ to RaFD	P_1	46.35	0.544
	P_2	48.18	0.386
	P_1 & P_2	**51.24**	**0.570**

The bold data indicates the best cross-database recognition performance

reference value for the understanding of facial expression behavior. Compared to other methods, the accuracy scores of the proposed method are 0.53%, 2.28% improvement. [1] employs HOG features for training binary CNNs and thereafter ensemble them to detect expressions. [35] utilizes surf features and applies a boosting algorithm to train classifiers. These two methods both build N networks for N expressions respectively, and have high and complex computations. Finally, the confusion matrix of the proposed method is presented in Fig. 7. We can observe that the recognition accuracies of the six basic expressions all exceed 90%.

4.7 Evaluation of cross-datasets performance

Facial expression and AU recognition methods still have problems achieving high accuracies and scores when evaluated using the cross-database validation protocol. Because of culture and race, different persons have different combinations of different AUs. Even though the environment is controlled within the database, the facial behaviors are not controlled within the database. Therefore, it is important to know the performance obtained by the model when it is trained by one database and tested over another database. As summarized in Table 8, with the help of prior knowledge named P_1 and P_2, the accuracies and F1-scores of expression and AU recognition are increased, indicating that the mapping relationship between expressions and AUs has a positive impact on the generalization performance. However, overall, the cross-dataset performance is much worse than the within-dataset performance. This may be attributed to the dependency of prior knowledge on the limited two databases. Specifically, the RaFD database labels the six basic emotion images strictly according to the specific AU combination. However, the CK+ database includes prototypes and major variants of each emotion, which means that the sequences are collected from a looser condition. Because of different data sources, the prior knowledge has a different distribution; more databases are required to improve generalization. Additionally, the recognition results of the model trained by RaFD database are both higher than the model trained by CK+ database. The sample sizes of RaFD are nearly four times those of CK+ database, indicating that the generalization of the model is limited by the size of the training sets.

5 Conclusion

In this paper, we present a novel approach for recognizing expressions and AUs, which is based on FE-AURDGCN. First, a generative model is trained by cGAN to filter identity information and extract expression information. Thereafter, we consider the dependency among AUs to construct an expression graph and embed the nodes with multiple AU-related patches extracted from the generative model. Finally, we use prior knowledge matrices to represent the strong dependencies between expressions and AUs and subsequently integrate them into the loss function to constrain the model. Experimental results on the extensively used CK+ and RaFD datasets have demonstrated the superiority of the introduced framework over the state-of-the-art methods. In the future, we plan to explore how to combine the temporal information of the sequences into network to improve performance.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 62077009).

References

1. Ali G, Iqbal MA, Choi TS (2016) Boosted NNE collections for multicultural facial expression recognition. Pattern Recogn 55:14–27. https://doi.org/10.1016/j.patcog.2016.01.032
2. Aquaro V, Bardoscia M, Bellotti R, Consiglio A, Carlo FD, Ferri G (2010) A Bayesian networks approach to operational risk. Phys A Stat Mech Appl 389(8):1721–1728
3. Cai J, Meng Z, Khan A, Li Z, O'Reilly J, Tong Y (2019) Identity-free facial expression recognition using conditional generative adversarial network. arxiv:1903.08051
4. Dapogny A, Bailly K, Dubuisson S (2018) Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. Int J Comput Vision 126(2):255–271. https://doi.org/10.1007/s11263-017-1010-1
5. Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. Proc Natl Acad Sci U S A 111(15):E1454. https://doi.org/10.1073/pnas.1322355111
6. Ekman P (1979) About brows: emotional and conversational signals. Cambridge University Press, Cambridge
7. Ekman P, Friesen W (1978) The facial action coding system: a technique for measurement of facial movement. Consulting Psychologists Press, Palo Alto
8. Ekman P, Friesen W, Hager J (2002) Facial action coding system: the manual on CD-ROM. Salt Lake City
9. Hao L, Wang S, Peng G, Ji Q (2018) Facial action unit recognition augmented by their dependencies. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG), pp 187–194. https://doi.org/10.1109/FG.2018.00036

10. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B (2014) Generative adversarial nets. In: Neural information processing systems, pp 2672–2680
11. Jun H, Xiaocui Y, Lejun Y (2019) Facial emotion and action unit recognition based on bayesian network. In: In 8th international conference on computing and pattern recognition (ICCPR), vol 1187, pp 1–7. https://doi.org/10.1117/12.2539053
12. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE international conference on computer vision (ICCV), pp 2983–2991. https://doi.org/10.1109/ICCV.2015.341
13. Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition? In: IEEE international conference on computer vision workshop (ICCVW), pp 19–27. https://doi.org/10.1109/ICCVW.2015.12
14. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: 5th international conference on learning representations (ICLR)
15. Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD (2011) Adversarial machine learning. In: ACM workshop on security and artificial intelligence, pp 43–58
16. Lai Y, Lai S (2018) Emotion-preserving representation learning via generative adversarial network for multiview facial expression recognition. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG), pp 263–270
17. Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg AD (2010) Presentation and validation of the Radboud faces database. Cognit Emotion 24(8):1377–88
18. Li G, Zhu X, Zeng Y, Wang Q, Lin L (2019) Semantic relationships guided representation learning for facial action unit recognition. In: The thirty-third AAAI conference on artificial intelligence, pp 8594–8601. https://doi.org/10.1609/aaai.v33i01.33018594
19. Li S, Deng W (2018) Deep facial expression recognition: a survey. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.2981446
20. Li Y, Wang S, Zhao Y, Ji Q (2013) Simultaneous facial feature tracking and facial expression recognition. IEEE Trans Image Process 22(7):2559–2573
21. Libralon GL, Romeio RA (2013) Inverstigating facial features for identification of emotions. In: IEEE conference on robotics and biomimetics (ROBIO), pp 1294–1299
22. Liu M, Li S, Shan S, Chen X (2013) Au-aware deep networks for facial expression recognition. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp 1–6. https://doi.org/10.1109/FG.2013.6553734
23. Liu M, Shan S, Wang R, Chen X (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE conference on computer vision and pattern recognition, pp 1749–1756. https://doi.org/10.1109/CVPR.2014.226
24. Liu Y, Zhang X, Lin Y, Wang H (2019) Facial expression recognition via deep action units graph network based on psychological mechanism. IEEE Trans Cognit Dev Syst. https://doi.org/10.1109/TCDS.2019.2917711
25. Tran L, X., Liu X (2017) Disent angled representation learning GAN for pose-invariant face recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1415–1424
26. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn–Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition—workshops, pp 94–101
27. Mahmoud Khademi LPM (2014) Relative facial action unit detection. In: IEEE conference on computer vision and pattern recognition (CVPR)
28. Maja P, Michel V, Ron R, Ludo M (2005) Web-based database for facial expression analysis. In: IEEE international conference on multimedia & expo. https://doi.org/10.1109/ICME.2005.1521424
29. Martín Abadi AA, Barham P, et al (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. CoRR. arxiv:1603.04467
30. Mengyi L, Shaoxin L, Shiguang S, Ruiping W, Xilin C (2015) Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian conference on computer vision, pp 143–157
31. Ming-Yu Liu OT (2016) Coupled generative adversarial networks. In: 29th conference on neural information processing systems (NIPS)
32. Mirza M, Osindero S (2014) Conditional generative adversarial nets. CoRR. arxiv:1411.1784
33. Ning Sun Qi Li RHJLGH (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61. https://doi.org/10.1016/j.patterc.2017.10.022
34. Pons G, Masip D (2018) Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. CoRR. arxiv:1802.06664
35. Rao Q, Qu X, Mao Q, Zhan Y (2015) Multi-pose facial expression recognition based on surf boosting. In: IEEE conference on affective computing and intelligent interaction (ACII), pp 630–635
36. Lu S, Dou Z, XJJN, Wen J (2019) Psgan: A minimax game for personalized search with limited and noisy click data. In: ACM SIGIR conference on research and development in information retrieval, pp 555–564
37. Karras T, Laine S, Aila T (2018) A style-based generator architecture for generative adversarial networks. In: IEEE conference on computer vision and pattern recognition (CVPR)
38. Velusamy S, Kannan H, Anand B, Sharma A, Navathe B (2011) A method to infer emotions from facial action units. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2028–2031. https://doi.org/10.1109/ICASSP.2011.5946910
39. Wang Z, Li Y, Wang S, Ji Q (2013) Capturing global semantic relationships for facial action unit recognition. In: 2013 IEEE international conference on computer vision, pp 3304–3311
40. Zhou Y, Shi BE (2017) Action unit selective feature maps in deep networks for facial expression recognition, pp 2031–2038
41. Tong Y, Liao W, Ji Q (2006) Inferring facial action units with causal relations. IEEE Comput Soc Conf Comput Vis Pattern Recognit (CVPR) 2:1623–1630
42. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 2168–2177. https://doi.org/10.1109/CVPR.2018.00231
43. Yu C, Xie C, Zhang Y (2015) A facial expression recognition strategy based on template matching. In: International congress on image and signal processing (CISP), pp 835–840. https://doi.org/10.1109/CISP.2015.7407993
44. Zeiler MD, Fergus R (2013) Visualizing and understanding convolutional networks. CoRR. arxiv:1311.2901
45. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503. https://doi.org/10.1109/LSP.2016.2603342

46. Zhao K, Chu W, De la Torre F, Cohn J, Zhang H (2016) Joint patch and multi-label learning for facial action unit and holistic expression recognition. IEEE Trans Image Process 25(8):3931–3946

47. Zhu Y, Wang S, Yue L, Ji Q (2014) Multiple-facial action unit recognition by shared feature learning and semantic relation modeling. In: 2014 22nd international conference on pattern recognition, pp 1663–1668. https://doi.org/10.1109/ICPR.2014.293