Knowledge-Based Systems 251 (2022) 109304

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Four-player GroupGAN for weak expression recognition via latent expression magnification

Wenjia Niu [a,b], Kaihao Zhang [c], Dongxu Li [a], Wenhan Luo [d,*]

[a] College of Engineering and Computer Science, Australian National University, Canberra, 2601, Australia
[b] School of electronic and information engineering, Hebei University of Technology, Tianjin, 300401, China
[c] University of Technology Sydney, Sydney, NSW 2007, Australia
[d] Sun Yat-sen University, Guangzhou, 510275, China

ARTICLE INFO

Article history:
Received 13 March 2022
Received in revised form 24 May 2022
Accepted 17 June 2022
Available online 22 June 2022

Keywords:
Face expression recognition
Weak expression
GAN
Deep CNN
Four players

ABSTRACT

Facial expression recognition has a wide range of applications in the real world. Although many existing deep learning methods have achieved remarkable success, *weak expression recognition* remains a challenging task because of the significant domain gap between a weak expression and its peak expression counterpart. One idea to solve this problem is to find an effective way to bridge the gap between the two domains by either transfer learning or cross-domain image synthesis. In this paper, we propose a Group Generative Adversarial Network (GroupGAN) that recognizes weak facial expression by magnifying the expressions to stronger or peak ones. Different from the traditional GAN which typically has only one generator and one discriminator, the proposed GroupGAN has one generator, one extractor and two discriminators. Similar to the "two-player game" analogy of the traditional GAN, in our setting the generator along with feature extractor act as one group to compete with the other group of the two distinct discriminators. Extensive experiments show that the proposed GroupGAN significantly improves the performance of weak expression recognition, and is able to magnify weak expressions, thus facilitating many expression-related vision tasks like sketch recognition.

1. Introduction

Facial expression recognition [1–5], which aims to predict facial expression labels like anger, disgust, fear, happiness, sadness and surprise, has attracted considerable attention from the computer vision community. It has a wide range of applications in fields such as human–computer interaction [6], face alignment [7] and medical diagnosis [8]. Traditional facial expression recognition methods primarily target at recognizing peak expressions (see Fig. 1(a) bottom row). While they achieved satisfactory performance on those peak expressions, most of them are unable to work on cases where the expressions are weak or subtle, as shown in Fig. 1(a) second row. Formally, *weak expressions* are referred to all the intermediate states between a neutral face and its peak expression. Judged by visual appearances, weak expressions are different from their peak counterparts, yet existing facial expression recognition systems are often trained on peak data, thus unable to distinguish weak expressions satisfactorily. Moreover, since many weak expressions are subtle, even similar to neutral faces, directly re-train a learning system on weak data will not lead to the desired recognition performance.

In this work, we propose a novel idea to tackle such a weak expression recognition problem. Instead of focusing on learning a better discriminative classifier, we consider a generative alternative. Specifically, we realize that, if one is able to transfer the image of a weak facial expression to a peak one while preserving the class label of the expression as well as the face identify, the task of weak-expression recognition can be conducted in the peak version, which is much easier to solve. We implement this idea by adopting the Generative Adversarial Networks (GAN) framework.

The traditional GAN consists of a single Discriminator (D) and a single Generator (G). One possible way of applying a traditional GAN for the task of transferring a weak-expression to its peak expression counterpart is to feed the weak-expression image to the generator G as a condition, and let G generate an image, intending to pass the "real or fake" testing conducted by D. Unfortunately, apart from this "real or fake" decision we have no way to effectively enforce that the generated images must be the realistic peak counterpart. Variants of GAN such as Wasserstein GAN [9] and CycleGAN [10] did not solve this issue neither. Recently, SinGAN [11] is proposed to manipulate images, which inspires us to address the weak expression recognition with multiple generators and discriminators.

* Corresponding author.
E-mail address: whluo.china@gmail.com (W. Luo).

https://doi.org/10.1016/j.knosys.2022.109304
0950-7051/

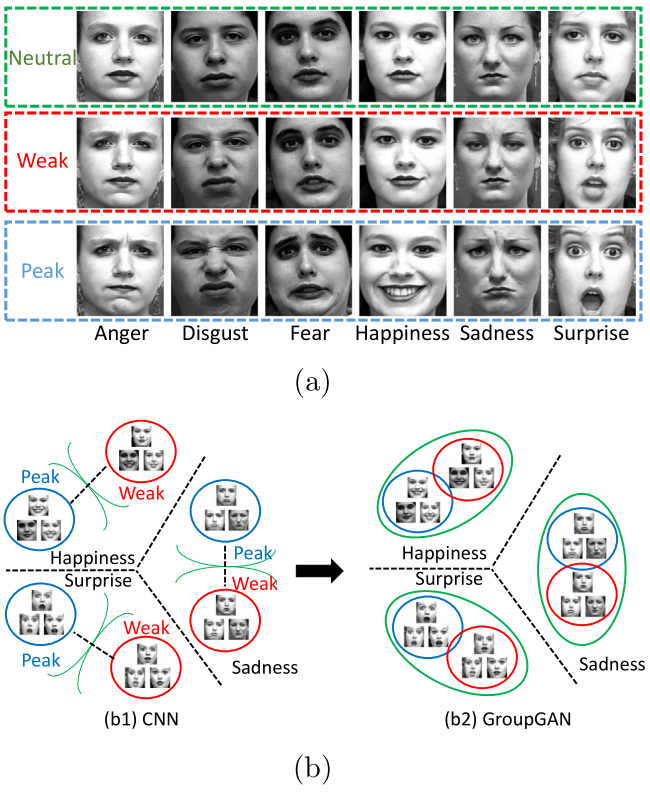

Fig. 1. (a) Examples of neutral, weak, and peak facial expressions; (b) Motivation of our model. The red and blue ellipses represent the features of weak and peak expressions, respectively. (b1) shows the features of latent space learned by traditional CNN. Faces with the same expression label may be classified into two different expression classes. (b2) shows the features learned by our GroupGAN model. We require features learned from one expression class must be close in feature space.

Specially, to address this problem, we realize that since there are multiple properties to expect, and a single discriminator is unable to perform multiple tasks, we therefore develop a GroupGAN which has multiple distinct discriminators–each of which takes care of only one aspect of the desired properties. Together they solve the multi-task problem. Specifically, we have two goals (or two desired properties) to achieve: (1) the generated peak expression and the corresponding real peak expression must have similar feature embeddings; (2) the generated peak expression must share the same expression class label with the weak input.

In order to achieve the first goal, we introduce a second discriminator to a GAN, denoted as D', where D' and D work together as one group (or team) jointly to fulfill the task. We call this scheme the "`1 Versus 2`" ($1V2$) scheme. Particularly, D aims to distinguish real images from fake images like conventional GAN, while D' distinguishes two domains with respect to expression conditions (peak or weak). G attempts to improve image generation quality in order to fool D and D'. The network converges when both D and D' are unable to tell between "`real or fake`" and "`peak or weak`", respectively. Note however, our second goal above requires that the generator G is able to not only synthesize a realistically looking peak expression, but also can distinguish different expression class labels, otherwise it would not be able to fulfill the second goal, *i.e.* retaining the class label. To this end, we further introduce an "Extractor" (denoted as E) to be a buddy of G. This new scheme is called "`2 Versus 2`" ($2V2$) GroupGAN. In particular, it consists of one Extractor (E) and one Generator (G) as one group, and two Discriminators (D and D') as the competing group (*i.e.* the opponent player). E extracts expression features.

Intuitively, during training G attempts to generate more realistic peak facial expression images, which can fool both D (whose task is to tell apart "fake" or "real") and D' (whose task is to determine "weak" or "peak"), while at the same time the mission of E is to extract more expressive CNN features such that to make the life of D' harder (*i.e.*, make it more difficult for D' to judge whether the features come from peak expression, or a weak one). A concept illustration of the expected output of GroupGAN is shown in Fig. 1(b). Traditional CNN may learn expression features with large intra-class variations (see b1), while features learned from one expression by our GroupGAN are all similar (see b2).

We conducted extensive experiments on three standard and commonly used datasets: CK+, Oulu-CASIA and LSEMSW, showing that the proposed GroupGAN model is able to learn powerful expression features and achieves superior performance than traditional models. Moreover, intermediate outputs of our G network are realistic and meaningful "magnified" version of the input weak facial expression, which can be used for other novel applications. Last but not least, our new two-versus-two groupGAN architecture can be applicable in many other vision tasks where multiple competing goals are expected simultaneously.

In summary, our main contribution is three-fold:

- We propose a GroupGAN framework which includes an extractor, a generator, and two different discriminators. It extends the competition between generator and discriminator in traditional GAN to two groups.
- We simultaneously handle weak facial expression recognition and emotion exaggeration aiming to exaggerate weak expression to a peak expression. In our proposed GroupGAN, the conventional discriminator together with an additional discriminator compete with the generator and extractor, synthesizing facial images that are of high quality and preserve expression type.
- Experimental results demonstrate that our proposed method outperforms the previous best methods in weak facial expression recognition on widely used facial expression databases.

2. Related work

2.1. Deep facial expression recognition

Deep learning has been adopted to solve facial expression recognition task [12]. For example, a deformable part learning component is incorporated into a 3D-CNN framework to capture the expression information in [13]. Yu and Zhang [14] propose a method with multiple deep CNN models. Each of them is initialized randomly and pre-trained to learn expression features. A deeper CNN with inception structures is employed in [15] for automated facial expression recognition. Zhao et al. [16] present a peak-piloted deep network and a back-propagation procedure called peak gradient suppression for learning expression features. Hu et al. [17] develop a multi-task learning method to jointly detect facial landmark and address expression recognition. Some two-stream CNN models are proposed by [18] and [19] to recognize facial expression via combining different kinds of features. To solve the issue that the facial expressions are not aligned with the pre-defined semantic categories, Vemulapalli and Agarwala [20] propose to learn a so-called compact expression embedding (16-dimension) by annotating facial expression labels in a fashion of a triplet. The derived compact embedding demonstrates its effectiveness in recognizing facial expressions. An end-to-end Compositional Generative Adversarial Network (Comp-GAN) is developed by Wang et al. in [21] to generate desired facial images of specific expressions and poses as data augmentation. In specific, there are two components in CompGAN, one for changing

the pose and the other one for changing the expression. Identity is maintained in the procedure of generation. The facial images produced by Comp-GAN benefit the following recognition task, which is verified by extensive experiments. In [22], both pose variation and identity bias are tackled by adversarial feature learning. With features learned from an encoder, two discriminators classifying pose and subject respectively are trained in an adversarial manner with the encoder. By doing so, this method is robust to pose variation and subject bias. Pan et al. [23] address a challenging problem, *i.e.* recognizing the expression of occluded facial image. The core idea is employing non-occluded facial image as privileged information to guide the recognition of occluded facial expression. The guidance is applied in both label space and feature space, and the guidance is carried out in an adversarial manner. Similarly, in [24], thermal images are used to improve the facial expression recognition of visible facial images. The thermal images play as a supplementary role by enhancing the visible features with adversarial learning and similarity constraints. They also enhance the ability of the classifier for the visible facial expression images. Experiments on public dataset show this fusion is effective. Rather than paying attention to facial expression recognition in the controlled environments, Zhang et al. [25] focus on recognizing facial expression in the wild. By modeling the problem as a domain adaptation problem, a Cycle-consistent adversarial Attention Transfer model (CycleAT) is proposed to jointly generate and recognize facial images. Cycle-consistency and adversarial loss are enforced in the proposed model, which are proved effective by experiments. Chen et al. [26] leverage the topological information of the labels from related distinct tasks like action unit recognition and facial landmark detection for facial expression recognition. Wei et al. [27] and Potamias *et al.* [28] make use of web data and 3D information to learn better emotion representation, respectively. Wang et al. [29] and Zhan et al. [30] study the problem of uncertainties and zero-shot for facial expression recognition, respectively.

2.2. Generative adversarial networks

GAN [31] has been successfully applied to face related vision applications. Many variants of GAN have been developed. Mirza et al. [32] propose a conditional GAN model which includes a conditional label to generator and discriminator to control the process of generation. Arjovsky et al. [9] introduce a new algorithm named Wasserstein GAN to improve the stability of learning and get rid of problems like mode collapse. Radford et al. [33] combine the traditional CNN and GAN model to propose a DCGAN framework. GAN is used for super-resolving digital images [34], removing motion blurs [35,36], image dehazing [37,38]. Meanwhile, it is also widely applied in text to image translation [39,40] and video prediction [41,42].

Almost all of them follow the paradigm that one generator synthesizes fake images and one discriminator tries to predict the reality of synthesized images. Recently, a SinGAN is proposed to manipulate/generate images to address multiple tasks. Different generators and discriminators are trained under the "coarse-to-fine" scheme to generate different results. SinGAN achieves great success, which inspires us to recognize weak expression. In our task, in order to exaggerate facial expression and reduce the intra-class variations, we propose a new GAN framework called GroupGAN which includes one extractor, one generator and two discriminators, to recognize weak expression via magnifying them.

3. GroupGAN

Our GroupGAN is designed based on traditional GAN. In this section, we first present our basic $1V2$ GroupGAN model for enhancing emotion features. Then, our full $2V2$ GroupGAN model is introduced. Finally, we conduct an analysis of the GroupGAN. In the following, we use lowercase letters to denote scalar or index, capital letters to denote a function approximator or process, bold letters for tensors.

3.1. Exaggeration features: 1V2

Traditional GAN consists of a generator G and a discriminator D. Given a facial image $\mathbf{X}^{weak} \in \mathbb{R}^{W\times H}$, we feed it into the generator as a condition. Through a stack of layers, the generator outputs an image of peak expression $\mathbf{Y}^{peak} \in \mathbb{R}^{W\times H}$, denoted as $\mathbf{Y} = G(\mathbf{X})$. To ensure that the content of the generated images resembles realistic images, we use MSE loss to measure the similarity between a pair of images, as defined:

$$\mathcal{L}_{content}^{MSE} = \left\| \mathbf{X}^{peak} - G(\mathbf{X}^{weak}) \right\|_2^2 , \tag{1}$$

where $\mathbf{X}^{peak}$ is the real-world peak expression image, and $G(\mathbf{X}^{weak})$ corresponds to a synthesized image which is generated from generator G. Moreover, the perceptual loss is also popularly used. However, in our study, our experimental results show that both of them achieve similar performance. Therefore, during the training stage, we employ MSE loss function to help update the proposed GroupGAN.

The discriminator D takes both real and fake images as input and aims to separate them. The learning process is with an adversarial loss defined as,

$$\mathcal{L}_{adversarial} = \log(1 - D(G(\mathbf{X}^{weak}))) . \tag{2}$$

The overall loss function for network training is a weighted combination of the above terms:

$$\mathcal{L} = \mathcal{L}_{content} + \lambda \cdot \mathcal{L}_{adversarial} , \tag{3}$$

where $\mathcal{L}_{content}$ can be Eq. (1), λ is a hyper-parameter to balance the content and adversarial losses.

As discussed previously, a single D is unable to fulfill both requirements. As a remedy, we introduce a second discriminator D', leading to our basic GroupGAN (1V2) framework.

Fig. 2 illustrates the difference between a conventional GAN and our basic GroupGAN. The basic GroupGAN model has three modules, G, D and D'. The main difference is that our basic GroupGAN has an additional discriminator D', which aims to determine whether the features obtained are from peak expression or weak expression. The new loss is defined as:

$$\min_D V_D = \Phi_D(\mathbf{X}_{real}) + \phi_D(\mathbf{X}_{fake}) , \tag{4}$$

$$\min_{D'} V_{D'} = \Phi_{D'}(\mathbf{X}^{peak}) + \phi_{D'}(\mathbf{X}_{fake}) , \tag{5}$$

$$\begin{aligned} \min_G V_G = {} & \Phi_G(\mathbf{X}_{real}, \mathbf{X}_{fake}) + \alpha_G \cdot \phi_D(\mathbf{X}_{fake}) \\ & + \beta_G \cdot \phi_{D'}(\mathbf{X}_{fake}) + \gamma_G \cdot \Phi_C(\mathbf{X}_{fake}) , \end{aligned} \tag{6}$$

where Φ_D plus ϕ_D, $\Phi_{D'}$ plus $\phi_{D'}$, Φ_G and Φ_C represent the energy functions of D, D', G and C. α_G, β_G and γ_G are the hyper-parameters to balance different loss functions.

D aims to distinguish real images from fake ones. This is done by minimizing $\Phi_D(\mathbf{X}_{real})$ and $\phi_D(\mathbf{X}_{fake})$. D' is trained to distinguish peak expression from weak ones, conditioning on the input images. G is trained to fulfill three requirements. Firstly, it should minimize the L_2 distance between real and fake images, defined by $\Phi_G(\mathbf{X}_{real}, \mathbf{X}_{fake})$. Secondly, G competes with D and D' in order to produce more realistic peak expression that minimizes

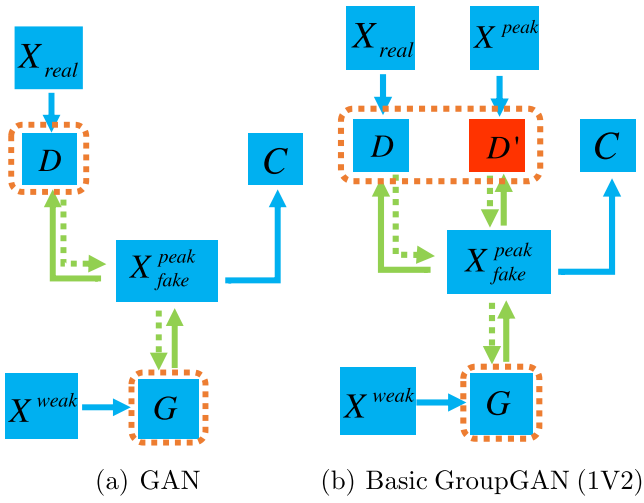

Fig. 2. (a) is the conventional GAN framework for expression exaggeration. It is composed of a generator and a discriminator. A real image $\mathbf{X}^{weak}$ is fed into the generator and the output is a generated image $\mathbf{X}^{peak}$. D is a discriminator to distinguish real and synthesized images. C is a facial expression classifier. (b) is the basic GroupGAN (1V2) framework. The blue arrows show the forward-only data flow and the green arrows show the adversarial process between G and D models. It is clear that there are three models competing in our basic GroupGAN, while the conventional GAN has only two adversarial models.

$\phi_D(\mathbf{X}_{fake})$ and $\phi'_D(\mathbf{X}_{fake})$. Thirdly, the cross-entropy loss predicted by a classifier C must be minimized. C is used to ensure the stabilization of expression recognition, but does not participate in the competing process. Details of these modules in our basic GroupGAN framework are elaborated on in the sequel.

Discriminator D. The discriminate network is designed to distinguish whether the input images are real or not and constrain the generated images to be similar to real face images. D is a binary classifier which is trained with the loss as,

$$\Phi_D(\mathbf{X}_{real}) = -\log(D(\mathbf{X}_{real})), \tag{7}$$

$$\begin{aligned}\phi_D(\mathbf{X}_{fake}) &= -\log(1 - D(\mathbf{X}_{fake})) \\ &= -\log(1 - D(G(\mathbf{X}^{weak}))),\end{aligned} \tag{8}$$

where $\log(1 - D(G(\mathbf{X}^{weak})))$ is the probability that the generated image is a real image. The overall loss function for D is shown in Eq. (4).

Discriminator D'. The second discriminator, D', solves a two-class classification problem and its loss function in Eq. (5) is,

$$\Phi_{D'}(\mathbf{X}^{peak}) = -\log(D'(\mathbf{X}^{peak})) \quad , \tag{9}$$

$$\begin{aligned}\phi_{D'}(\mathbf{X}_{fake}) &= -\log(1 - D'(\mathbf{X}_{fake})) \\ &= -\log(1 - D'(G(\mathbf{X}^{weak}))),\end{aligned} \tag{10}$$

where $\log(1 - D'(G(\mathbf{X}^{weak})))$ is the probability that the generated image is a peak expression.

Classifier C. To ensure the expression of synthesized images is same to the input images, C learns expression features of the real and fake images. C discriminates these expression by classifying them using c different labels, and thus C can be regarded as a solver of a c-class classification problem. The loss function of C is defined using the cross-entropy loss.

$$\Phi_C(X_{fake}) = -\sum P(\mathbf{X}_{fake}) \log Q(\mathbf{X}_{fake}) \ , \tag{11}$$

where $\mathbf{X}_{fake}$ corresponds to the generated image. $P(\mathbf{X}_{fake})$ is the true distribution of facial expression and $Q(\mathbf{X}_{fake})$ is the predicted distribution.

Generator G. Eq. (6) shows that the loss function of G consists of four kinds of energy functions. Except ϕ_D, ϕ'_D and Φ_C, which have been introduced above, G is also trained by minimizing $L2$ distance between real and fake images, defined as Φ_G as,

$$\Phi_G(\mathbf{X}_{real}, \mathbf{X}_{fake}) = \left\|\mathbf{X}_{real} - \mathbf{X}_{fake}\right\|_2^2 . \tag{12}$$

Two difficulties arise though. First, the generator has to maintain the correct label of expression during the generation process. As a result it would have to take care of two tasks at the same time: (1) generating realistic image to recognize and retain the class label of the input expression. It is too much from a single model to jointly recognize and retain labels and generate images; (2) the generator has to compete with two discriminators. Specially, the generator needs not only fool the first discriminator such that it cannot distinguish real or fake, but also fool the second discriminator whose mission is to tell apart peak from weak.

3.2. Separation learning: 2V2

To overcome the above difficulties, we introduce an additional extractor E into the basic GroupGAN to help the generator compete with the two discriminators. Conventional GAN generates images by deconvolutional layers, and thus does not make use of valuable expression information during generation. In order to obtain the expression information, the extractor takes the facial images as input and captures the semantic expression features. Rather than concatenating the original facial images and the expression vectors directly, in our GroupGAN, we employ an encoder–decoder architecture as our generator and utilize the middle-level features captured by the generator to incorporate with the emotion vectors. Specifically, at the bottleneck of the generator, the features learned by the generator are concatenated with the emotion vector. As Fig. 3 shows, the blue and green cubes represent features captured by encoder and the extractor E, respectively. After combining the two kinds of features, the decoder up-samples them to the size of input images. The loss functions for D, D' and G are the same as Eqs. (4), (5) and (6), while the loss functions of E is defined as,

$$\min_E V_E = \Phi_C(\mathbf{X}_{fake}) + \alpha_E \cdot \phi_D(\mathbf{X}_{fake}) + \beta_E \cdot \phi_{D'}(\mathbf{X}_{fake}), \tag{13}$$

where α_E, β_E are hyper-parameters to balance different energy functions. The energy functions of D and D' are replaced by:

$$\Phi_D(\mathbf{X}_{real}) = -\log(D(\mathbf{X}_{real}, \mathbf{F}_{exp})), \tag{14}$$

$$\begin{aligned}\phi_D(\mathbf{X}_{fake}) &= -\log(1 - D(\mathbf{X}_{fake})) \\ &= -\log(1 - D(G(\mathbf{X}^{weak}, \mathbf{F}_{exp}))),\end{aligned} \tag{15}$$

$$\Phi_{D'}(\mathbf{X}^{peak}) = -\log(D'(\mathbf{X}^{peak}, \mathbf{F}_{exp})), \tag{16}$$

$$\begin{aligned}\phi_{D'}(\mathbf{X}_{fake}) &= -\log(1 - D(\mathbf{X}_{fake})) \\ &= -\log(1 - D(G(\mathbf{X}^{weak}, \mathbf{F}_{exp}))),\end{aligned} \tag{17}$$

Architecture. The generator G is an encoder–decoder structure. The encoder has four layers and each convolutional layer of encoder is followed by one residual block [43]. The sizes of filters are 7, 5, 3 and 3, respectively. The decoder includes four deconvolutional layers and two convolutional layers with filter size 3. Our two discriminators are VGG-like structures which are constructed by convolutional layers and fully connected layers. Facial images are fed into the network while the captured expression information is fed into the network as conditional information. With the help of the two discriminators, the expression features can be embedded into the generator and the extractor can extract more powerful expression features. The architecture of the extractor is

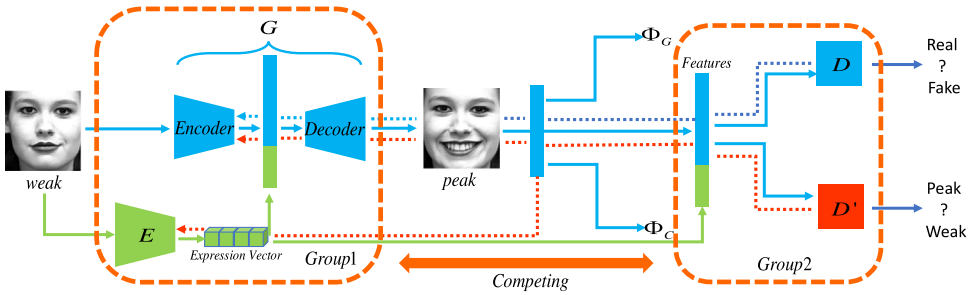

Fig. 3. The architecture of our GroupGAN framework (2V2). The framework consists of two groups. The first group has a generator and an extractor. The generator aims to generate facial images with peak expression, while the extractor aims to extract expression feature information to aid the process of generation. The second group has two different discriminators. One discriminator is to distinguish the real and fake images, while the other discriminator is to distinguish the peak and weak expression. The competing process between them can be explained as: the generator tries to synthesize facial images which can fool the two discriminators, and the extractor tries to extract expression features which make it difficult for the second discriminator (red) to distinguish whether the expression is peak or weak.

similar to classifier, but without the last fully-connected layer. In the experiments, we will replace the extractor and the classifier with popular CNN-based models to evaluate our GroupGAN. The iterative training process is shown in Algorithm 1.

3.3. Further analyses of GroupGAN

The goal of the generator in GroupGAN is to discover latent differences between peak expression $\mathcal{P}$ and weak expression $\mathcal{W}$, and then transfer input images from $\mathcal{W}$ to $\mathcal{P}$. The transformation process can be described as

$$\mathbf{p} \approx \pi\ (\mathbf{w}; \mathbf{e})\ , \tag{18}$$

where $\mathbf{w}$ and $\mathbf{p}$ are the instances from $\mathcal{W}$ and $\mathcal{P}$, respectively, $\mathbf{e}$ is the expression vector captured by the extractor, π is the approximator function that controls the transfer.

Perceptual loss can measure the distance between two kinds of images in high-level semantic space. Models trying to learn the transformation of expression information via deep structures can be represented as

$$\Psi(\mathbf{p}) \approx \Psi\ (\pi(\mathbf{w}; \mathbf{e}))\ , \tag{19}$$

where Ψ denotes the expression exaggerating process via deep module. Li et al. [44] shows that the transfer can be represented as linear shifting. Therefore, the difference between two domains can be formulated as

$$\Delta\mathbf{v}_{perceptual} = \|\Psi\ (\mathbf{p}) - \Psi\ (\pi(\mathbf{w}; \mathbf{e}))\ \|\ , \tag{20}$$

where $\Delta\mathbf{v}_{perceptual}$ is the discrepancy between peak and weak expression in high-level feature space. We will discuss the discrepancy in "Feature visualization" in the experiment section.

The content difference in pixel space can be formulated as,

$$\Delta\mathbf{v}_{content} = \|\mathbf{p} - \pi(\mathbf{w}; \mathbf{e})\|\ . \tag{21}$$

To show what the GroupGAN focuses at, we can compute $\Delta\mathbf{v}_{content}$ with the form of a heat map as:

$$\mathbf{H}_{i,j} = \sum_{k} \Delta\mathbf{v}^2_{i,j,k}\ , \tag{22}$$

where i, j, k denote the width, height and channel indexes, respectively. The heat map of shifting will be discussed in the experiment section.

4. Experiments

4.1. Databases

The CK+ Database. CK+ is an extended version of Cohn–Kanade (CK) [45]. It consists of 123 subjects with 593 video sequences,

Algorithm 1 The training procedure of GroupGAN

Initialize the parameters of G, E, D and D' as θ_G, θ_E, θ_D and $\theta_{D'}$. Define learning rates of different parts as ρ_G, ρ_E, ρ_D and $\rho_{D'}$. Set the hyper-parameters of different energy functions as α_G, β_G, γ_G, α_E and β_E.
Pre-train E **on peak expression sets.**
Pre-train the encoder and decoder on facial images.
for # of iterations **do**
Update D**:**
Sample n real facial images $\mathbf{X}_{real}$ as a batch from the training set.
Synthesize n fake facial images $\mathbf{X}_{fake}$ by G.
$\theta_D := \theta_D + \rho_D \nabla_{\theta_D} \frac{1}{n} \sum_{i=1}^{n} (\log D(\mathbf{X}^i_{real}) + \log(1 - D(\mathbf{X}^i_{fake})))$
Update D'**:**
Sample n peak expression images $\mathbf{X}^{peak}$ as a batch from the training set.
Synthesize n fake facial images $\mathbf{X}_{fake}$ by G.
$\theta_{D'} := \theta_{D'} + \rho_{D'} \nabla_{\theta_{D'}} \frac{1}{n} \sum_{i=1}^{n} (\log D'(\mathbf{X}^{peak}_i) + \log(1 - D'(\mathbf{X}^i_{fake})))$
Update G**:**
Sample n real facial images $\mathbf{X}_{real}$ as a batch from the training set.
Synthesize n fake facial images $\mathbf{X}_{fake}$ by G.
$\theta_G := \theta_G - \rho_G \nabla_{\theta_G} \frac{1}{n} \sum_{i=1}^{n} (\left\|\mathbf{X}^i_{real} - \mathbf{X}^i_{fake}\right\|^2_2 + \alpha_G \cdot \log(1 - D(\mathbf{X}^i_{fake}))$
$+ \beta_G \cdot \log(1 - D'(\mathbf{X}^i_{fake})) + \gamma_G \cdot P(\mathbf{X}^i_{fake}) \log(Q(\mathbf{X}^i_{fake})))$
Update E**:**
Sample n real facial images $\mathbf{X}_{real}$ as a batch from the training set.
Synthesize n fake facial images $\mathbf{X}_{fake}$ by G.
$\theta_E := \theta_E - \rho_E \nabla_{\theta_E} \frac{1}{n} \sum_{i=1}^{n} (p(\mathbf{X}^i_{fake}) \log(q(\mathbf{X}^i_{fake})) + \alpha_E \cdot \log(1 - D(\mathbf{X}^i_{fake}) + \beta_E \cdot \log(1 - D'(\mathbf{X}^i_{fake}))$
end for

which are labeled as seven types of emotion, including anger, contempt, disgust, fear, happiness, sadness, and surprise. The document of this database shows that their expression video sequence begins from the neutral expression and ends with the peak expression. Similar to [16], we denote the 6-*th* to 8-*th* frames as "weak expressions" as they show non-peak expression with very weak intensities. The last frame of each sequence is regarded as the peak expression. We divide the sequences into 5 groups and employ a 5-fold validation protocol. In each time, four groups are selected as training sequences and the rest one is for testing.

The Oulu-CASIA Database. Oulu-CASIA includes 80 subjects of 23 to 58 years old with six expressions (anger, disgust, fear, happiness, sadness and surprise), and these facial expressions are captured under three different illumination conditions, including normal, weak and dark. There are 80 subjects with six expressions, thus 480 sequences in total under each illumination condition. All these video sequences represent expression from neutral condition to peak condition. We evaluate our model under the popular normal illumination condition. Similar to the

CK+ database, we also denote the 6-*th* to 8-*th* frames as "weak expressions" based on the document of the Oulu-CASIA database. A 5-fold validation protocol is employed.

The LSEMSW Database. Different from the above laboratory-controlled databases, the LSEMSW database [17] is collected from more than 200 movies and TV serials. It consists of 175,679 facial images, which are cropped by MTCNN [46]. It is much more challenging for two reasons. Firstly, it includes thirteen kinds of subtle expression (happy, anxious, sad, scared, angry, hesitant, indifferent, surprised, arrogant, thinking, helpless, suspicious, and questioning), which is finer than CK+ and Oulu-CASIA datasets. Secondly, all expressions are subtle or weak in the wild. This dataset is divided into three sets for training (80%), validation (10%) and testing (10%) [17]. This database does not include pairs of peak and weak expression. To adapt to our method, we first train a CNN model on training set and predict on the validation set. Then we choose facial images which are easily predicted by our pre-trained CNN model as face images of peak expression.

4.2. Implementation details

When training GroupGAN, we utilize the Gaussian distribution with zero mean and a standard deviation of 0.01 to initialize weights of the generator, extractor and discriminators. We update all weights after learning a mini-batch of size 4 in each iteration. To augment the training set, 128 $\times$ 128 patches are cropped at random locations in resized images (144 $\times$ 144). We also randomly flip frames (horizontally) in the training stage. The generator is trained with a learning rate from 10^{-4} and we decrease the learning rate to 10^{-5} when the training loss does not decrease. The biases are initialized as 1. In all layers, the momentum is set as 0.9 and the weight decay is set as 0.005. Hyper-parameters for adversarial learning are set to 2×10^{-3}.

4.3. Weak facial expression recognition

To show the performance of weak facial expression recognition and demonstrate the effectiveness of the proposed GroupGAN model, we apply six popular BaseNet models, including AlexNet [47], VGG11 [48], VGG13 [48], VGG16 [48], VGG19 [48], ResNet18 [43], ResNet34 [43] and ResNet50 [43], into our framework. All the BaseNets are representative methods for learning high-level representation to recognize images. The difference from the original versions is that we reduce the neuron number of full-connected layers from 4096 to 512 or 256. Based on these BaseNets, we develop three competing frameworks, *CNN*, *DCGAN*, *GroupGAN(1V2)* and *GroupGAN(2V2)* to conduct ablation analyses.

- **CNN** is a popular network mentioned above. The input of this model is weak expression facial images. The CNN extracts features from the convolutional layers and fully-connected layer. Finally, the representation will be input to a c-class classifier to predict the expression label. The model is trained based on the loss function in Eq. (11).
- **DCGAN** is a popular two-player network [49], which includes a generator and a discriminator. The input of this model is weak expression facial images, and the output is synthesized peak expression. Finally, the synthesized images will be input to a classifier to predict the expression label.
- **GroupGAN(1V2)** is our proposed basic GroupGAN including a generator and two discriminators. This version of GroupGAN model is updated based on the loss functions in Eqs. (4), (5) and (6). The energy functions in these functions are in Eqs. (7), (8), (9), (10), (11) and (12).

Table 1
Performance comparison on the mix of CK+ and Oulu-CASIA databases in terms of average classification accuracy (%) by the 5-fold cross-validation. The best results are shown in bold, which also applies to the following tables.

BaseNet	CNN	DCGAN	GroupGAN (1v2)	**GroupGAN (2v2)**
AlexNet	62.61	63.29	63.74	**67.12**
VGG11	64.64	65.77	66.22	**68.02**
VGG13	64.86	66.67	67.34	**70.05**
VGG16	63.74	64.86	65.55	**68.69**
VGG19	61.26	62.61	63.74	**64.19**
ResNet18	59.91	60.14	60.81	**62.61**
ResNet34	58.58	59.47	59.92	**62.21**
ResNet50	57.85	58.32	58.25	**61.89**

Table 2
Confusion matrix on the mix of CK+ and Oulu-CASIA databases in terms of average classification accuracy (%) by the 5-fold cross-validation.

	An	Di	Fe	Ha	Sa	Su
An	**71.4**	10.1	8.6	2.7	5.4	1.8
Di	17.4	**62.1**	4.2	5.2	9.3	1.8
Fe	4.5	5.4	**65.6**	6.9	15.2	2.4
Ha	16.3	2.4	3.6	**67.4**	3.9	6.4
Sa	15.5	10.3	5.2	3.3	**58.5**	7.2
Su	4.2	8.2	2.4	12.6	3.3	**69.3**

- **GroupGAN(2V2)** is our proposed GroupGAN including a generator, an extractor and two discriminators. The loss functions employed to train this model are Eqs. (4), (5), (6) and (13). The energy functions in these functions are in Eqs. (11), (12), (14), (15), (16) and (17).

The effectiveness of GroupGAN. To show the effect of GroupGAN, we mix the training and testing sets of the CK+ and Oulu-CASIA datasets together and compare the performance of the above three different methods on them. Results are shown in Table 1. we observe that, (1) GroupGAN(1V2) achieves higher accuracies than CNN and DCGAN methods. This verifies the effectiveness of the basic GroupGAN method by introducing an additional discriminator D'. As shown in Fig. 2, the D' along with D as a group competes with generator to push the generative model to synthesize images of less difference from real images with peak expression. (2) The improvement of GroupGAN(2V2) from GroupGAN(1V2) reveals the advantage of the additional extractor, which learns to capture expression information as a conditional vector to reduce the learning stress of the generator and guide it to synthesize peak-emotion facial images. (3) In addition, we can find that the model based on the VGG13 achieves the best performance. The reason might be that the expression recognition is not as challenging as the ImageNet competition. Therefore, we can choose a moderately sized backbone like VGG13. Moreover, as shown in Table 2, the confusion matrix of different types of expression is represented. It is observed that, the expression of angry, happy, and surprise is easier to be recognized. We suspect that these three types of expression exhibit evident differences from other ones when exaggerated.

Feature Visualization. We are also curious about what the GroupGAN has learned, thus we analyze the feature maps by this model. Fig. 4 shows the final 256-dimension feature maps of the proposed GroupGAN model and traditional CNN model, respectively. Feature differences are normalized to a range of 0$\sim$1. Locations will be lighted on if their values are larger than a fixed threshold value. Thus fewer light-on positions correspond to the more similar features of facial images. These examples show that features of an identical expression learned by GroupGAN have more commonly activated neurons than the features captured by traditional CNN models.

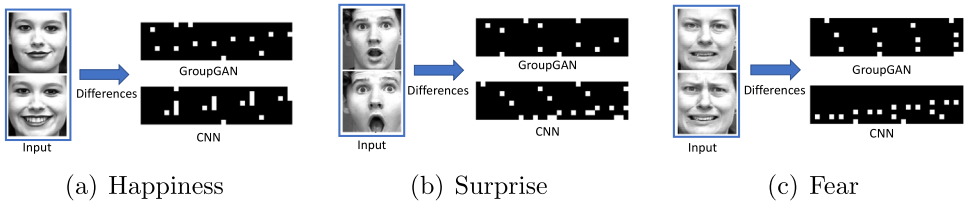

Fig. 4. The difference of the captured 256-dimension expression features between two facial images. The fewer light-on positions, the more similar their features are. The bottom rows in the right correspond to features learned based on CNN model, while the top rows are features learned based on GroupGAN. We rearrange the features as 8×32 for the sake of convenient illustration.

Table 3
Performance comparison on the LSEMSW database in terms of average classification accuracy (%).

Method	Accuracy
LPQ	10.86
LBP	10.53
EOH	13.47
AlexNet	26.77
VGG	28.07
ResNet18	33.47
RN+LAF+ADA [17]	36.72
SCN [29]	36.85
GroupGAN	**37.17**

Comparing with State-of-the-art Methods. We compare the proposed method with the state-of-the-art methods in Table 3. Previous approaches for expression recognition use hand-crafted features, like LPQ [50], LBP [51] and EOH [52]. EOH considers both spatial and texture information, while LBP and LPQ primarily learn texture information. We find that EOH achieves better performance than the other two kinds of hand-crafted features. However, all the three kinds of features achieve bad performance because weak expression recognition is a very challenging problem. Recently, popular deep CNNs like AlexNet [47], VGGNet [48] and ResNet [43] are employed for emotion recognition [53]. SCN [29] is a recent state-of-the-art method, which is also employed in our comparison. As Table 3 shows that these deep learning methods achieve better performance. The previous best method is RN+LAF+ADA [17], which achieves the performance based on a multi-task learning strategy and under the help of facial landmark information. For the LSEMSW database, our designed GroupGAN replaces pixel-wise loss with perceptual loss which is calculated by extractor. Table 3 shows that GroupGAN achieves satisfactory performance which outperforms the state-of-the-art method.

4.4. Expression magnification

Finally we study the intermediate results of our method to show whether our GroupGAN can improve the performance of weak expression recognition. Given a facial image with weak expression, the extractor learns their expression features and the encoder of G encodes their attribution information. These two types of information are fused and fed into the decoder of G to synthesize a facial image with peak expression. Some examples are shown in Fig. 5. The first column displays the neutral faces. The third, and fifth columns are the generated faces with magnified facial expression, and real peak expression, respectively.

GroupGAN automatically learns most informative features for expression magnification. To visualize it, we plot heat maps (see Eq. (22)) in Fig. 5. The second and fourth columns are heat maps of the generated, and real peak expression images, respectively.

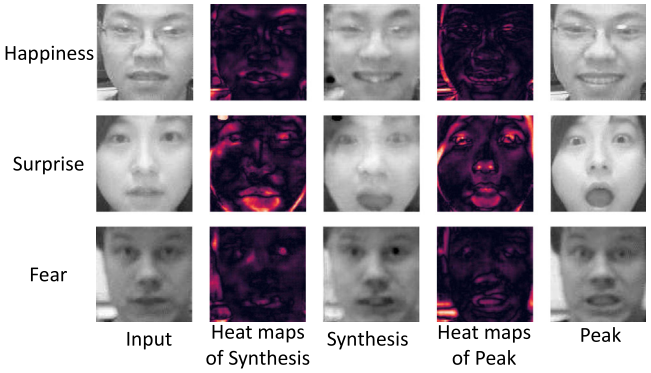

Fig. 5. Results of emotion magnification for different types of expression. From left to right, they are input facial images, synthesized facial expressions and their heat maps, real peak expressions and their heat maps, respectively. Obviously, it is easier to recognize the facial expression based on the synthesized results than input.

It is evident that our model clearly captures the intuition that the most expressive facial regions are mouth, nose, eyes or eyebrows, *i.e.*, consistent with human perception. Fig. 5 shows that our proposed GroupGAN improves the performance based on enhancing the facial expression and focuses on important facial organs.

4.5. Rethinking and discussion of GroupGAN

Other ways. To distinguish peak expression from weak ones, the proposed GroupGAN trains another D'. We do not use the likelihood of an expression classifier for the same purpose (for example, the input images have peak expression if the likelihood are high.) because the likelihood of a classifier for peak and non-peak expression are both high in some easily recognizable expression, such as happiness. However, it is easier for D' to distinguish.

Expression labels. The $\mathbf{F}_{exp}$ is extracted by our E, which includes expression information as we mentioned above. To guarantee that the extracted feature from E carries expression information, we rely on a classifier via $\Phi_C(X_{fake})$ in Eq. (13). Table 1 shows the effectiveness.

The difference between D and D'. Apart from the difference of learning goals, the training samples between them are also different. The positive samples for D are all real peak and fake expression, while the positive samples for D' are only real peak expression.

4.6. The GroupGAN for other tasks

For weak expression recognition, the proposed GroupGAN improve the classification accuracy via transferring weak expression to their peak version. In fact, apart from weak expression recognition, it can be also applied for other tasks. For example, hand-drawn sketch recognition is a fundamental issue in

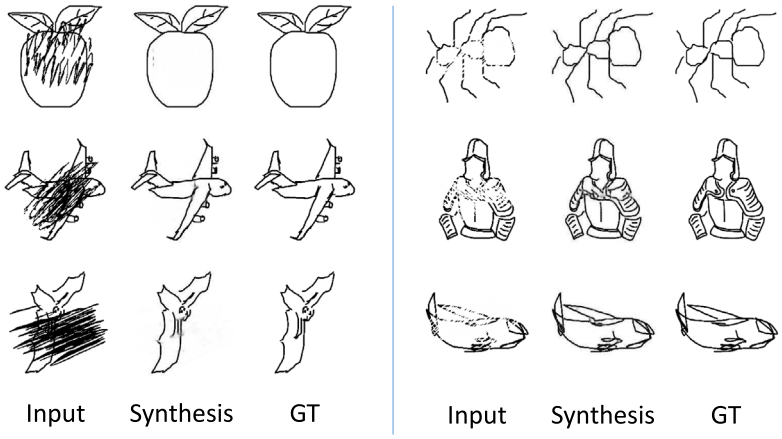

Fig. 6. Results of sketch modification for different input. From left to right, they are input sketch images, synthesized sketch images and their ground-truth sketch images, respectively. The GroupGAN can remove the unwanted parts and add the missing information.

computer vision. A complete sketch is easy to be recognized via some well-designed CNN models [54–56]. However, it is difficult for them to recognize some incomplete or destroyed sketch images which usually miss several important information. The proposed GroupGAN can improve the accuracy of recognition via transferring the incomplete or destroyed sketch to complete and high-quality sketch images. To show the effectiveness of the proposed GroupGAN, we choose 1000 sketch images from the Sketchy Database [57], and synthesize destroyed versions via adding slashes to them. In this way, we can obtain pairs of perfect and destroyed sketch images. Finally, we train the proposed GroupGAN on them and evaluate the trained model on testing images. Fig. 6 shows the qualitative results of our GroupGAN to restore input sketches. From the left to right are the input sketch, the output of our GroupGAN and the ground-truth sketches, respectively. It is obvious that our GroupGAN can improve the quality of sketch images, which is beneficial to recognize the destroyed sketch images.

5. Conclusion

In this paper, we have presented a Group Generative Adversarial Network framework for weak facial expression recognition and emotion enhancing. The framework includes one generator, one extractor and two discriminators. The generator collaborates with the extractor as a group to compete with the two discriminators, which maximize their classification accuracy, whereas the generator and extractor reduce accuracy of the discriminators by synthesizing images of high quality with peak-like expression. Experimental results demonstrate that our framework not only improves the performance of popular CNN models but also outperforms state-of-the-art methods on weak expression recognition. In addition, our GroupGAN can enhance emotion via transferring from weak to peak expression.

CRediT authorship contribution statement

Wenjia Niu: Conceptualization, Methodology, Software, Data curation. **Kaihao Zhang:** Conceptualization, Methodology, Software, Data curation. **Dongxu Li:** Writing – original draft, Revision. **Wenhan Luo:** Writing – original draft, Revision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded in part by the ARC Centre of Excellence for Robotics Vision (CE140100016), ARC-Discovery (DP 190102261) and ARC-LIEF (190100080) grants, as well as a research grant from Baidu on autonomous driving. The authors gratefully acknowledge the GPUs donated by NVIDIA Corporation. We thank all anonymous reviewers and editors for their constructive comments.

References

[1] T. Ma, W. Tian, Y. Xie, Multi-level knowledge distillation for low-resolution object detection and facial expression recognition, Knowl.-Based Syst. (2022) 108136.
[2] F. Nan, W. Jing, F. Tian, J. Zhang, K.-M. Chao, Z. Hong, Q. Zheng, Feature super-resolution based facial expression recognition for multi-scale low-resolution images, Knowl.-Based Syst. 236 (2022) 107678.
[3] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, L. Xie, A novel feature separation model exchange-GAN for facial expression recognition, Knowl.-Based Syst. 204 (2020) 106217.
[4] Z. Sun, R. Chiong, Z.-p. Hu, Self-adaptive feature learning based on a priori knowledge for facial expression recognition, Knowl.-Based Syst. 204 (2020) 106124.
[5] H. Li, H. Xu, Deep reinforcement learning for robust emotional classification in facial expression recognition, Knowl.-Based Syst. 204 (2020) 106172.
[6] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain, Image Vis. Comput. (2009).
[7] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: European Conference on Computer Vision, ECCV, 2014.
[8] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, K.M. Prkachin, Automatically detecting pain using facial actions, in: International Conference on Affective Computing and Intelligent Interaction, ACII, 2009.
[9] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: The International Conference on Machine Learning, ICML, 2017.
[10] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: The International Conference on Computer Vision, ICCV, 2017.
[11] T.R. Shaham, T. Dekel, T. Michaeli, SinGAN: Learning a generative model from a single natural image, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2019.
[12] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Trans. Affect. Comput. (2020).
[13] M. Liu, S. Li, S. Shan, R. Wang, X. Chen, Deeply learning deformable facial action parts model for dynamic expression analysis, in: Asian Conference on Computer Vision, ACCV, 2014.
[14] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: The ACM on International Conference on Multimodal Interaction, ICMI, 2015.
[15] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: The Winter Conference on Applications of Computer Vision, WACV, 2016.

[16] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, in: European Conference on Computer Vision, ECCV, 2016.
[17] G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson, et al., Deep multi-task learning to recognise subtle facial expressions of mental states, in: European Conference on Computer Vision, ECCV, 2018.
[18] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: The International Conference on Computer Vision, ICCV, 2015.
[19] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. (TIP) (2017).
[20] R. Vemulapalli, A. Agarwala, A compact embedding for facial expression similarity, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[21] W. Wang, Q. Sun, Y. Fu, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y.-G. Jiang, X. Xue, Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression, in: Proceedings of the 27th ACM International Conference on Multimedia, in: MM '19, 2019, pp. 211–219.
[22] C. Wang, S. Wang, G. Liang, Identity- and pose-robust facial expression recognition through adversarial feature learning, in: Proceedings of the 27th ACM International Conference on Multimedia, in: MM '19, 2019, pp. 238–246.
[23] B. Pan, S. Wang, B. Xia, Occluded facial expression recognition enhanced through privileged information, in: Proceedings of the 27th ACM International Conference on Multimedia, in: MM '19, 2019, pp. 566–573.
[24] B. Pan, S. Wang, Facial expression recognition enhanced by thermal images through adversarial learning, in: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, ACM, 2018, pp. 1346–1353.
[25] F. Zhang, T. Zhang, Q. Mao, L. Duan, C. Xu, Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach, in: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, ACM, 2018, pp. 126–135.
[26] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui, Label distribution learning on auxiliary label space graphs for facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13984–13993.
[27] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, D. Samaras, Learning visual emotion representations from web data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13106–13115.
[28] R.A. Potamias, J. Zheng, S. Ploumpis, G. Bouritsas, E. Ververas, S. Zafeiriou, Learning to generate customized dynamic 3D facial expressions, in: European Conference on Computer Vision, Springer, 2020, pp. 278–294.
[29] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897–6906.
[30] C. Zhan, D. She, S. Zhao, M.-M. Cheng, J. Yang, Zero-shot emotion recognition via affective structural embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1151–1160.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: The Conference on Neural Information Processing Systems, NeurIPS, 2014.
[32] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.
[33] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: The International Conference on Learning Representations, ICLR, 2016.
[34] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[35] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, DeblurGAN: Blind motion deblurring using conditional adversarial networks, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[36] S. Nah, T.H. Kim, K.M. Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[37] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, M.-H. Yang, Gated fusion network for single image dehazing, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[38] R. Li, J. Pan, Z. Li, J. Tang, Single image dehazing via conditional generative adversarial network, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[39] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: The International Conference on Learning Representations, ICLR, 2016.
[40] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[41] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, in: The Conference on Neural Information Processing System, NeurIPS, 2016.
[42] X. Liang, L. Lee, W. Dai, E.P. Xing, Dual motion GAN for future-flow embedded video prediction, in: The International Conference on Computer Vision, ICCV, 2017.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[44] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, M.-H. Yang, Universal style transfer via feature transforms, in: The Conference on Neural Information Processing Systems, NeurIPS, 2017.
[45] T. Kanade, Y. Tian, J.F. Cohn, Comprehensive database for facial expression analysis, in: The IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2000.
[46] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. (SPL) (2016).
[47] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: The Conference on Neural Information Processing Systems, NeurIPS, 2012.
[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: The International Conference on Learning Representations, ICLR, 2015.
[49] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: The International Conference on Learning Representations, ICLR, 2016.
[50] A. Dhall, A. Asthana, R. Goecke, T. Gedeon, Emotion recognition using PHOG and LPQ features, in: The IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2011.
[51] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image Vis. Comput. (IVC) (2009).
[52] H. Meng, B. Romera-Paredes, N. Bianchi-Berthouze, Emotion recognition by two view SVM_2K classifier on dynamic facial expression features, in: The IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2011.
[53] C. Ravat, S. Solanki, Facial expression recognition using convolutional neural networks, in: The International Conference on Vision, Image and Signal Processing, ICVISP, 2018.
[54] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, T.M. Hospedales, Sketch-a-net: A deep neural network that beats humans, Int. J. Comput. Vis. (IJCV) (2017).
[55] R.K. Sarvadevabhatla, J. Kundu, Enabling my robot to play pictionary: Recurrent neural networks for sketch recognition, in: Proceedings of the ACM International Conference on Multimedia, ACMMM, 2016.
[56] K. Zhang, W. Luo, L. Ma, H. Li, Cousin network guided sketch recognition via latent attribute warehouse, in: Proceedings of the Association for the Advancement of Artificial Intelligence, AAAI, 2019.
[57] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: Learning to retrieve badly drawn bunnies, ACM Trans. Graph. 35 (4) (2016) 1–12.