HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

GCANet: Geometry cues-aware facial expression recognition based on graph convolutional networks

Shutong Wang [a,b,1], Anran Zhao [c,1], Chenghang Lai [d,1], Qi Zhang [e,*], Duantengchuan Li [h], Yihua Gao [f], Liangshan Dong [g], Xiaoguang Wang [a]

[a] *School of Information Management, Wuhan University, Wuhan 430072, China*
[b] *National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China*
[c] *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China*
[d] *School of Computer Science, Fudan University, Shanghai 200438, China*
[e] *School of Information Management, Central China Normal University, Wuhan 430079, China*
[f] *Key Laboratory of Sports Engineering of General Administration of Sport of China, Wuhan Sports University, Wuhan 430079, China*
[g] *School of Physical Education, China University of Geosciences, Wuhan 430074, China*
[h] *School of Computer Science, Wuhan University, Wuhan 430072, China*

ARTICLE INFO

Article history:
Received 24 February 2023
Revised 30 May 2023
Accepted 30 May 2023
Available online 8 June 2023

Keywords:
Facial expression recognition
Graph convolutional network
Geometry cue
Uncertainty
Emotion label distribution learning

ABSTRACT

Facial expression recognition (FER) task in the wild is challenging due to some uncertainties, such as the ambiguity of facial expressions, subjective annotations, and low-quality facial images. A novel model for FER in-the-wild datasets is proposed in this study to solve these uncertainties. The overview of the proposed method is as follows. First, the facial images are grouped into high and low uncertainties by the pre-trained network. The graph convolutional network (GCN) framework is then used for the facial images with low uncertainty to obtain geometry cues, including the relationship among action units (AUs) and the implicit connection between AUs and expressions, which help predict the probability of the underlying emotional label. The emotion label distribution is produced by combining the predicted latent label probability and the given label. For the facial images with high uncertainty, *k*-nearest neighbor graphs are built to determine the *k* facial images in the low uncertainty group with the highest similarity to the given facial image. The emotion label distribution of the given image is then replaced by fusing the emotion label distribution based on the distances between the given image and its adjacent images. Finally, the constructed emotion label distribution facilitates training in a straightforward manner using a convolutional neural network framework to identify facial expressions. Experimental results on RAF-DB, FERPlus, AffectNet, and SFEW2.0 datasets demonstrate that the proposed method achieved superior performance compared to state-of-the-art approaches.

1. Introduction

Facial expression recognition (FER) is one of the most important components for human–computer interactive system (Yang et al., 2008; Zhang et al., 2021) and artificial intelligence (Ondras et al., 2020; Kuruvayil and Palaniswamy, 2022). Human facial expression (Tian et al., 2001; Ma et al., 2022), which is a critical carrier of emotion perception, is also a signal that conveys emotional state and intention (Pantic and Rothkrantz, 2000). As a result, accurate FER has remarkable real-world significance in a number of fields, such as immersive and interactive technologies under digital museums (Donadio et al., 2022; Benford et al., 2022), facial expression recognition of children with autism (Dong et al., 2021b; Dong et al., 2021a), contactless sport training monitor (Wu et al., 2017) and abnormal expression detection in physical exercise (Nayak et al., 2021; Chen et al., 2019). The athletes' emotional state during the game is detected through facial expressions and body language. However, FER has many challenges. On the one hand, the early

[*] Corresponding author.

E-mail addresses: shutongwang1221@gmail.com (S. Wang), anranzhao415@whu.edu.cn (A. Zhao), chlai21@m.fudan.edu.cn (C. Lai), qi.zhang.3519@gmail.com (Q. Zhang), dtclee1222@gmail.com (D. Li).

[1] These authors contributed equally to this work.

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101605
1319-1578/© 2023 The Author(s). Published by Elsevier B.V. on behalf of King Saud University.

facial expression recognition algorithms (Zhang et al., 2019; Xie et al., 2019; Majumder et al., 2018; Rodriguez et al., 2017) mainly input the original facial image directly into the deep model. Such an approach not only affects the performance of the expression recognition but also fails to effectively use prior knowledge of cognitive neuroscience as the theoretical basis. Most algorithms maximize the facial action coding system (FACS) to solve this problem (Ekman, 1993). Ekman designed this system as a taxonomy of human facial expressions. FACS is a comprehensive and anatomical system that defines facial action units (AUs). These units are described by a series of atomic facial muscle actions. While facial AUs capture local variations on human faces, facial expression categories depict facial behaviors globally (Vasanthi and Seetharaman, 2022; Revina and Emmanuel, 2021).

Furthermore, FACS can encode diverse human face deformations based on the combination of AUs (Friesen and Ekman, 1978), thus providing comprehensive emotional categories. In daily life, various changes in the local muscles of the face occur when humans produce expressions. Therefore, accurately exploring and obtaining the representation of AUs for FER is a critical problem in the real world.

On the other hand, facial images collected in emotion recognition competitions, such as FER2013 (Tang, 2013), as well as emotion recognition in the wild (EmotiW) (Kahou et al., 2013), provide relatively sufficient data from challenging real-world scenarios. Such collections implicitly facilitate the development of FER from laboratory-controlled to wild settings in recent years. Inconsistent and incorrect labels for real-world facial expression images widely exist because high-quality annotation is extremely challenging due to uncertainty (Wang et al., 2020b). These uncertainties are often caused by various reasons. For example, the annotators with different psychological knowledge and background might have different perceptions. Thus, the bias arises from the subjectivity of the annotation of emotion labels (Zeng et al., 2018). Moreover, complex emotions may lead to the presence of a variety of mixed expressions, in which each basic emotion has a different component in various expressions, especially for in-the-wild datasets. The low quality of the images due to occlusion of key parts, large pose changes, and low resolution also cause uncertainties. The samples are illustrated in Fig. 1. Overall, training models on FER datasets with uncertainty may result in overfitting on uncertain facial images and hinder model convergence.

A new and efficient method, namely geometry cues-aware based on graph convolutional network (GCANet), is proposed in this paper for FER in-the-wild datasets to address the aforementioned problems. The motivation principally comes from the following observations. Fig. 2 displays a comparison of various expressions with their AUs. From top to bottom, the expressions demonstrate surprise, happiness, fear, anger, disgust and sadness. Thus, geometry cues of facial expressions in the real-world environment are varied and do not follow a certain standard form. For example, when feeling surprised one may open mouth wide, open eyes wide and raise eyebrows, or only open mouth slightly, open eyes and close mouth. Furthermore, there are a number of the same combinations of AUs for different expressions. When raising eyebrows and opening his mouth, a man may be surprised, or happy, or fear, or even angry. Therefore, this study aims to effectively extract the potential co-occurrence relationships between AUs to represent expressions. The specific locations of AUs are illustrated in Fig. 3. Inspired by graph convolutional network (GCN) (Scarselli et al., 2008; Kipf and Welling, 2016), which is proposed to pass and aggregate information in the graph structure, exploring the geometry cues containing the relationships among AUs and the mapping relations between AUs and expressions is possible. GCN has been extensively utilized in image classification (Wang et al., 2018), semantic segmentation (Liang et al., 2017) and relational reasoning (Battaglia et al., 2016), due to their strong advantages in capturing the discriminative feature representations. In addition, GCNs can capture the intrinsic connections between each vertex node in the graph by learning an adjacency matrix. Moreover, label distribution learning has been effective in addressing annotation bias and label ambiguity by indicating the degree to which each label can depict an instance. Thus, this study mainly aims to construct emotion label distribution through GCNs to learn AU representation and then facilitate training using a CNN framework to perform expression recognition.

Overall, the main contributions of this study could be threefold as follows.

- The GCN framework based on AUs graph representation is adopted to obtain prior knowledge of geometry cues and construct the emotion label distribution for the first time to the best of our knowledge.
- *K*-nearest neighbor graphs are utilized for the facial images with high uncertainty to gain emotion label distribution. Such a distribution is obtained by fusing the emotion label distributions of adjacent images according to the distances.
- The proposed GCANet can effectively deal with the inconsistent labels, label noise, and reduce the impact of uncertainty. The experimental results indicate that the proposed method performs better than the existing methods.

The rest of this paper is organized as follows. The related work in the FER domain, including AUs-based network methods and label distribution learning approaches, is briefly reviewed in Section 2. The details of the proposed GCANet are firstly introduced

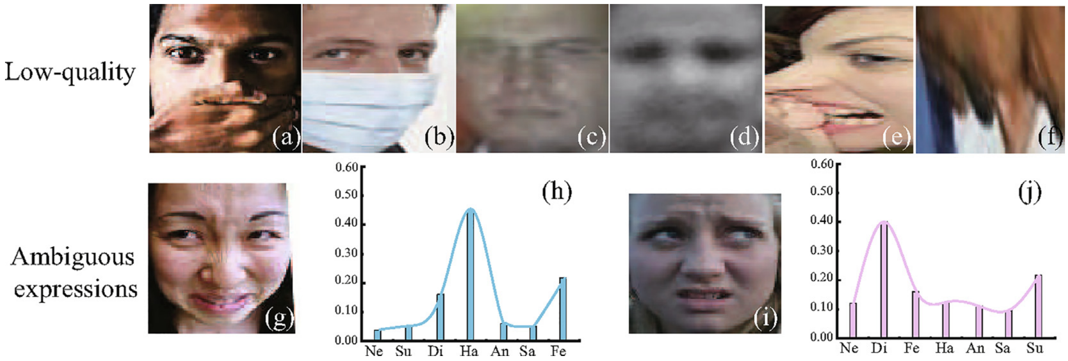

Fig. 1. Illustration of some real-world samples with high uncertainty from the from RAF-DB. (a) and (b) facial images with critical parts occlusions. (c) and (d) facial images with low resolution. (e) and (f) images with large pose-variant. (g) and (i) ambiguous facial images with their expression label distributions (h) and (j).

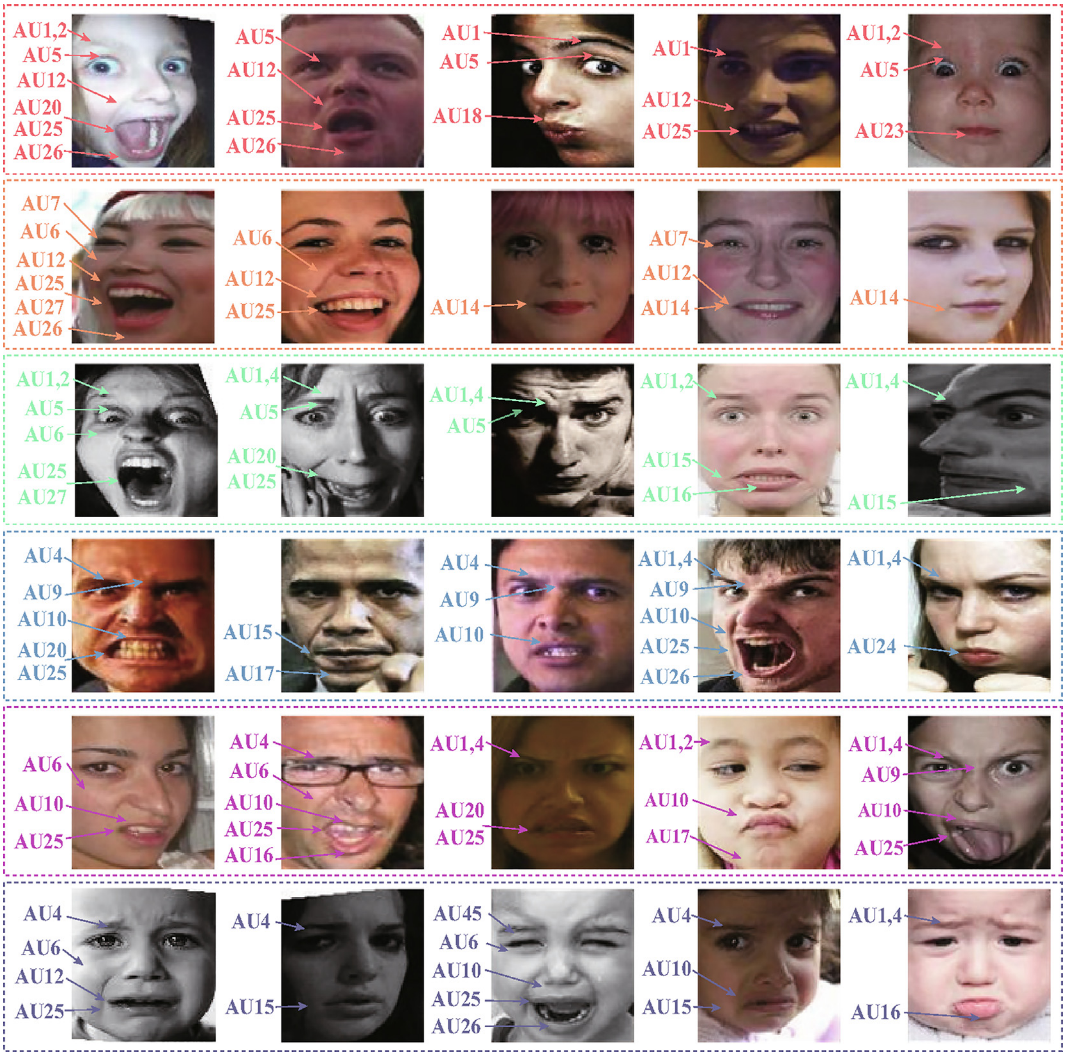

Fig. 2. Comparison of various expressions and their AUs from RAF-DB. From top to bottom: Surprise, Happiness, Fear, Anger, Disgust, Sadness.

in Section 3. The experimental settings, results, comparison, and analysis are provided in Section 4. Finally, the summary is presented in Section 5.

2. Related work

2.1. Facial expression recognition

Most of the FER methods have been recently proposed to identify several basic expressions in labeled facial images. The basic expressions are formally defined by Ekman (Ekman, 1993), and comprise happiness (Ha), anger (An), fear (Fe), sadness (Sa), disgust (Di), and surprise (Su). The automatic FER system consists of three fundamental components: human face detection, facial feature extraction, and facial expression recognition. In the first stage, an automatic face detector, such as MTCNN (Zhang et al., 2016), is utilized to position faces in complicated backgrounds. Then, the detected faces could be adopted to crop and align. In the next stage, the facial geometry and appearance features activated by an expression are captured from facial images. These features fall into two categories, namely engineered features (Eleftheriadis et al., 2014a; Zheng, 2014; Sangineto et al., 2014) and learning-based features (Liu et al., 2014; Eleftheriadis et al., 2014b; Gu et al., 2017), whether they are captured by hand-crafted descriptors or deep learning algorithms. The engineered features could comprise local features based on texture, while global features based on geometry. The hybrid feature combines two or more engineered features. Moreover, the learned features are learned from deep neural networks (Rifai et al., 2012). The winners (Tang, 2013; Kahou et al., 2013) of the FER2013 and Emotiw2013 challenge both extracted features by adopting deep CNNs. In (Rodriguez et al., 2017), the temporal appearance and geometry features are extracted by combining two different models simultaneously. In addition, the works of deep neural networks for FER tasks have been considerably increased owing to the emergence of large-scale training databases, such as RAF-DB (Li and Deng, 2018), AffectNet (Mollahosseini et al., 2017), and ExpW (Dhall et al., 2012). In (Li et al., 2018), Li et al. proposed a CNN framework with an attention mechanism to solve the occlusion problem by perceiving the occlusion parts of the face and focusing on the discriminative unblocked patches. A novel region attention network (Wang et al., 2020a) which can extract valuable regions of the face, is presented for FER with occlusion and pose diversity. In (Wang et al., 2020b), the self-cure network (SCN) model is developed to suppress the uncertainty of FER through the self-attention and careful relabeling mechanisms, thus achieving impressive results. Zheng et al. (Zheng et al., 2021) presented a two-stream spatial–temporal

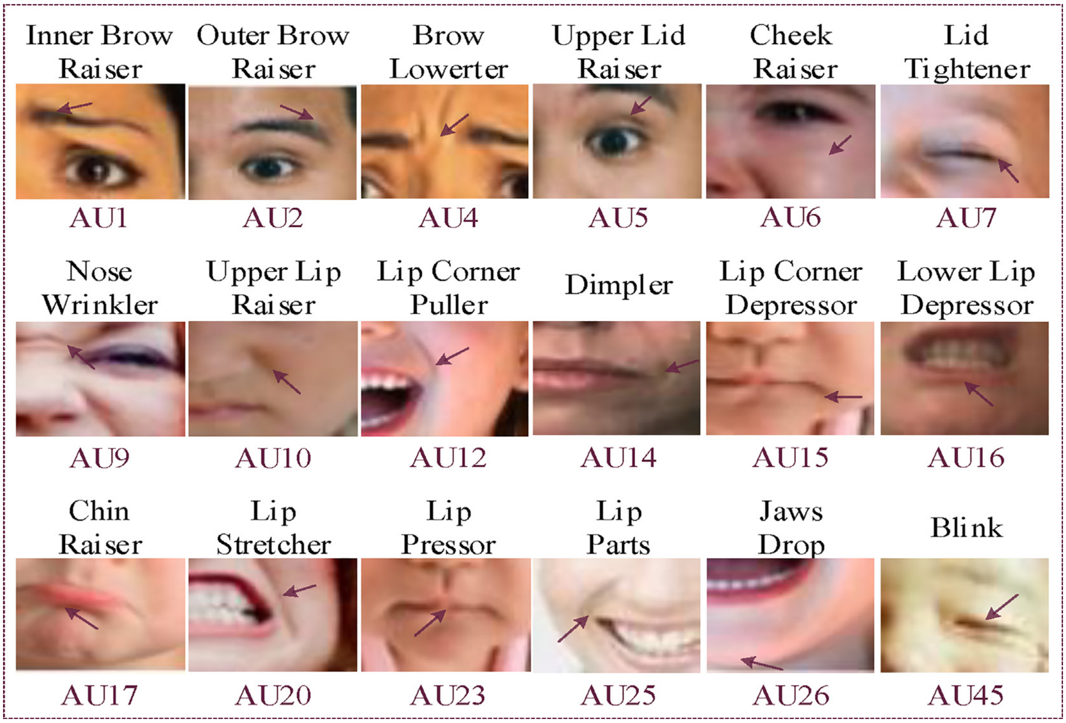

Fig. 3. Locations of action units on unconstrained faces.

network which introduce a Symmetry Loss and temporal module to extract the potential depth feature and multi-scale information respectively. To address the problem of insufficient samples, in (Zhao et al., 2022) put forward a hybrid-supervision learning framework that combines the advantages of supervised and unsupervised learning for expression recognition with fewer parameters.

2.2. Graph-based deep network

The variants of CNNs have been recently proposed to process graph data, especially in image and video classification tasks (Battaglia et al., 2018), due to their interpretability. The graph adjacency matrix is used as the main input form in the convolution process (Luo et al., 2017). Meanwhile, numerous methods adopting the graph model have been presented for FER and AU recognition tasks (Mohseni et al., 2014). In (Kung et al., 2015), a dual subspace nonnegative graph embedding, which is effective for the identity-independent problem of FER, is developed to represent facial expression images by employing identity and expression subspaces. In (Li et al., 2019), an innovative algorithm named graph-based dynamic ensemble pruning is proposed to tackle challenges in the classifier selection process in the dynamic integrated pruning methods. The dynamic ensemble pruning approach is considerably sensitive about the membership of the test sample neighborhood. Thus, a novel FER approach named deep AUs graph network, which is based on the psychological mechanism, is developed by Liu et al. (Liu et al., 2019). This network transforms the facial graph into a mutation adjacency matrix with a unique sequence. Moreover, global geometric and local appearance features are mixed by adopting a novel deep structure with sequential and diagonal convolutions, which works well for FER.

GCNs can effectively deal with graph data while maintaining the advantages of standard CNN, which are proposed as a new deep neural network learning framework. In addition, GCNs are suitable and powerful for diverse image classification tasks. In (Fan et al., 2020), the semantic correspondence convolution module is introduced to learn the semantic relationships among feature maps automatically, which enhances the discriminability of features.

2.3. Label distribution learning

The facial expression is not a single emotion category but usually blends with different basic emotions in Fig. 1(g) and (i), which would lead to the label ambiguity and incorrectness. Label distribution learning is one of the effective ways presented to mitigate the adverse impact of the aforementioned problem. This method indicates that the degree of each label could depict an instance. In (Li and Deng, 2019), a deep dual-manifold CNN framework, which could learn discriminative features by maintaining the manifold structure of expression labels and the local affinity of the deep features for multi-label expressions, is introduced. In (Zhou et al., 2015), a novel label distribution learning approach in auxiliary label space graphs is proposed to address annotation inconsistency. This approach assumes that facial images and their adjacent images should have a similar expression distribution in the label space of landmarks and AUs. In (Chen et al., 2020), a novel label distribution learning approach in auxiliary label space graphs is proposed to tackle the problem of annotation inconsistency, which assumes that facial images and their adjacent images should have similar expression distribution in the label space of landmarks and AUs. Zhang et al. (Zhang et al., 2020) suggested the correlation emotion label distribution learning model based on the similarity distributions of different expressions for infrared FER, which may not be fitted for large-scale real-world datasets.

Unlike their methods, the current study obtains geometry cues via GCN to construct the emotion label distribution and then facilitates learning in CNN for FER.

3. Proposed method

This section demonstrates the overall structure of the proposed GCANet as illustrated in Fig. 4. It consists of two primary stages

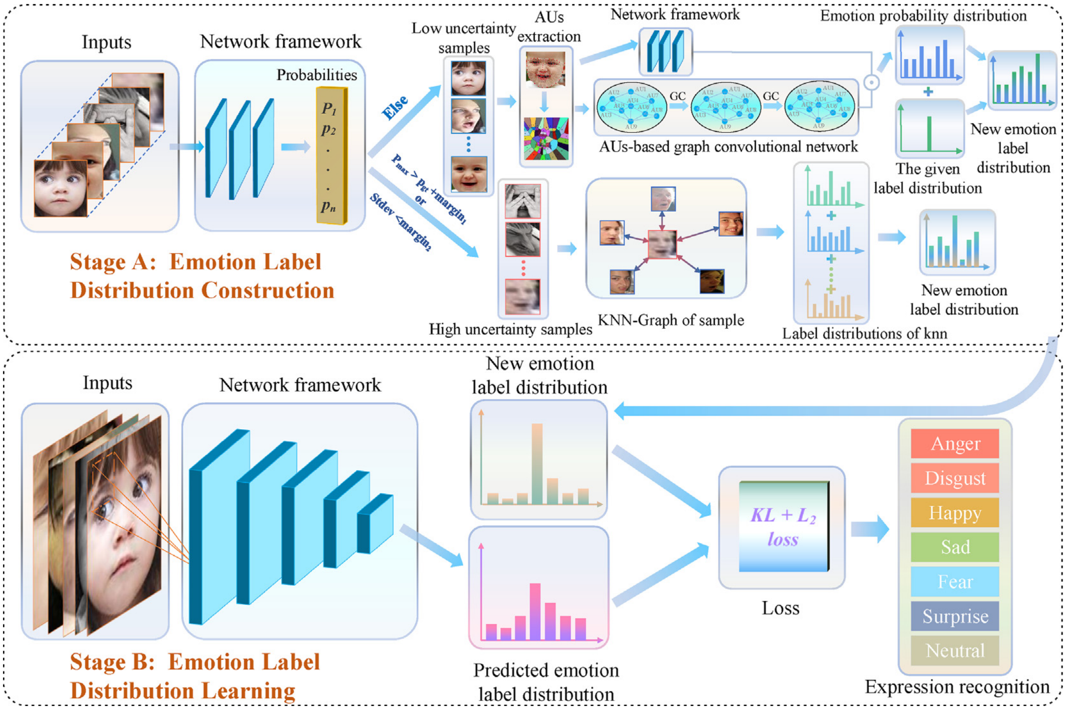
Inputs
Network framework
Probabilities
p_1
p_2
p_n
Stage A: Emotion Label Distribution Construction
Else
$p_{max} > p_{gt}$ +margin$_1$
or
Stdev <margin$_2$
Low uncertainty samples
AUs extraction
Network framework
Emotion probability distribution
GC
GC
AUs-based graph convolutional network
The given label distribution
New emotion label distribution
High uncertainty samples
KNN-Graph of sample
Label distributions of knn
New emotion label distribution
Inputs
Network framework
New emotion label distribution
Predicted emotion label distribution
KL + L$_2$ loss
Loss
Anger
Disgust
Happy
Sad
Fear
Surprise
Neutral
Expression recognition
Stage B: Emotion Label Distribution Learning

Fig. 4. Pipeline of the proposed GCANet model. Given a batch of facial images and the given labels, their emotion label distributions are produced in the stage A, the label distribution is then combined with our framework for optimization and training to achieve FER in the stage B.

which are emotion label distribution construction and emotion label distribution learning in the proposed GCANet. In the first stage, the given facial images would be divided into two sets of high and low uncertainty. For the low uncertainty samples, GCN is applied to investigate facial geometry cues. The new constructed emotion label distribution for the sample is then gained by fusing the emotion probability distribution learned from geometry cues and the given label to suppress uncertainty. However, it is difficult to accurately extract the feature of an AU which is occluded or a quite low-resolution region. Thus, the emotion probability distribution of the sample with high uncertainty is generated by the relationship among AUs, which may cause bias. The outputs of samples with high uncertainty have two main characteristics after the softmax function in the model. Firstly, the maximum prediction probability is greater than the probability of the given label, which indicates that the sample's expression is misclassified. The rule (i) is introduced to defined this, which is given as $p_{max} > p_{gt} + margin_1$, where p_{max} denotes the maximum prediction probability, p_{gt} stands for the probability of given label and $margin_1$ indicates a margin threshold. Secondly, there is not much difference in the probability of each expression, which indicates that the sample's expression is ambiguous and difficult to classify. The rule (ii) is used to describe this as $S_{tdev} < margin_2$, where S_{tdev} is the standard deviation of predicted expression probabilities and $margin_2$ represents a margin threshold. The approximate k-nearest neighbor (KNN) graphs are utilized for facial images with high uncertainty. The aim is to determine k facial images in the low uncertainty group that have the highest similarity to a given facial image. The emotion label distribution of the given image is subsequently replaced by fusing their emotion label distributions according to the distances between the given images and its adjacent images. From a global perspective, the images that share similar global AUs characteristics may have close emotion distribution. We choose integrating multiple expression image emotional distribution instead of the greatest similarity as an alternative to this image with high uncertainty, which could reduce the incidental bias. Thus, the emotion probability distribution generated by the KNN graph method is less biased than GCN. In the second stage, the new constructed emotion label distribution would be able to train directly in the traditional way by a CNN framework to complete the facial expression recognition tasks. More details are to be discussed in the subsequent subsections.

3.1. Emotion label distribution construction

3.1.1. AUGCN-based label distribution construction

AUs Graph Construction: Graph convolutional network is introduced to explore geometry cues. The node representations are then updated by propagating information between nodes. Let $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \in R^{n \times z(0)}$ represent the feature descriptions of AUs, where n indicates the number of AUs and z is the dimensionality of AUs features. The function $\sigma(\cdot)$ on a graph G is adopted to represent the GCN, which is different from the standard convolutions operating on local region in an image. Each GCN layer could be represented as

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right). \tag{1}$$

where $\mathbf{A} \in R^{n \times n}$ is referred to as the corresponding correlation matrix, $\sigma(\cdot)$ denotes an activation function which is LeakyReLU (Eleftheriadis et al., 2014a) in our experiments. $\mathbf{H}^{(l)} \in R^{n \times z(l)}$ denotes the matrix of activations in the l^{th} layer, $\mathbf{H}^{(0)} = \mathbf{U}, \mathbf{W}^{(l)} \in R^{n \times z(l+1)}$ means a layer-specific trainable weight matrix and $\mathbf{H}^{(l+1)} \in R^{n \times z(l+1)}$ indicates the output of activation in the $(l+1)^{th}$ layer. The complex inter-relationships of the nodes can be learned and modeled by stacking multiple GCN layers. The final output is $\mathbf{W}^{(L)} \in R^{z(L) \times c}$, where c represents the class of expressions.

Moreover, the correlation matrix plays a pivotal role in GCN which updates the representation of nodes through information propagation between nodes. Therefore, constructing the correlation matrix for the proposed model GCANet is crucial. In this paper, the correlation matrix $\mathbf{A}$ is built through a data-driven way. Specifically, the correlation among AUs is defined by exploring their co-occurrence patterns in the expression database. The correlation dependency of AUs is modeled as conditional probability, i.e., $p(AU_j|AU_i)$ which represents the probability of the occurrence of AU_j when the AU_i appears. It should be noted that $p(AU_j|AU_i)$ is not equal to $p(AU_i|AU_j)$. Then, the number of times that AUs occur in the training dataset needs to be counted in order to obtain the matrix $\mathbf{M} \in R^{n\times n}$, which is used for the correlation matrix construction. Specifically, n stands for the quantity of AUs, M_{ij} indicates the concurring times of AU_i and AU_j. The next, the conditional probability p_i is obtained via adopting AUs co-occurrence matrix, which is calculated by

$$p_i = M_i / N_i, \tag{2}$$

where M_i represents the times that AU_i occur in the given samples and N_i means the number of occurrences of AU_i in the training dataset.

The correlation matrix $\mathbf{A}$ can be computed by

$$A_{ij} = \begin{cases} p_{ij} / \sum_{j=1, i\neq j}^{n} p_{ij}, & i \neq j, \\ 1 - p_{ij}, & i = j. \end{cases} \tag{3}$$

where $p_{ij} = p(AU_j|AU_i)$ denotes the probability of AU_j when AU_i occurs. And AU_i stands for the category of AU, n indicates the number of $AU, n = 1, 2, \ldots, 45$.

Emotion Label Distribution Construction: Given an image $\mathbf{X}$, let $\mathbf{Y}$= $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_c\}$ indicate probability matrix of c possible basic expressions, and $\mathbf{D}_i = \{\mathbf{d}_{i1}, \mathbf{d}_{i2}, \cdots, \mathbf{d}_{ic}\}$ denote the emotion label distribution related to $\mathbf{X}_i$. All CNN base frameworks could be applied to capture the features from the facial images. Following the works of (Chen et al., 2020; Bargal et al., 2016; Yang et al., 2018), the ResNet-50 (He et al., 2016) is adopted to extract facial features as the skeleton network in our experiments. The image $\mathbf{X}$ is input with the 224×224 pixels. Subsequently, the feature vector of the "conv5_x" layer from neural network is obtained, whose size is $2,048 \times 7 \times 7$. Then, the max-pooling and average-pooling operations are combined as the hybrid pooling layer. The hybrid pooling layer is utilized to obtain the image-level feature $\mathbf{V}$, which is defined as

$$\mathbf{V} = f_{HYP}(f_{\text{cnn}}(\mathbf{X}; \boldsymbol{\theta}_{cnn})) \in R^D, \tag{4}$$

where $\boldsymbol{\theta}_{cnn}$ represents model parameters, f_{cnn} denotes the feature vector of the "conv5_x" layer from neural network, f_{HYP} indicates the feature vector of the hybrid pooling layer, and $\mathbf{D}$ is set to 2,048. In this way, the inter-dependent emotion probability distribution is learnt. According to the learned image feature representations $\mathbf{V}$ as well as the obtained weight matrix $\mathbf{W}$, the predicted emotion probability distribution $\hat{\mathbf{y}}_i$ can be expressed by

$$\hat{\mathbf{y}}_i = \mathbf{W}^{\mathrm{T}}\mathbf{V}. \tag{5}$$

Moreover, we set the given label distribution of a facial image $\mathbf{X}_i$ as $\mathbf{y}_i \in R^c$, where y_{ij} denotes the probability of each emotion appearance and is in the range [0, 1]. The constructed emotion label distribution should conform to a principle that the sum of its element needs to be equal to 1. Hence the normalized label distribution as,

$$g_{ij} = \frac{\exp(\hat{y}_{ij})}{\sum_k \exp(\hat{y}_{ik})}. \tag{6}$$

The given labels of datasets are considered in the proposed method due to their credibility. Thus, the new constructed emotion label distribution is defined as

$$d_{ij} = \varepsilon g_{ij} + (1 - \varepsilon) y_{ij}, \tag{7}$$

where g_{ij} represents the probability of emotion label distribution generated by the GCN, y_{ij} is the probability of the given label distribution of a facial image and ε denotes the weight parameter balanced the constructed emotion label distribution and the given label distribution. In addition, our method requires only traditional cross-entropy loss function to train the entire emotion label distribution construction network, with the following equation

$$L = -\frac{1}{c}\sum_j y_{ij} \ln \hat{y}_{ij}. \tag{8}$$

where $c$denotes the classes of expressions, y_{ij} and $\hat{y}_{ij}$ represent the ground truth and predicted label respectively.

3.1.2. KNN graphs-based label distribution construction

The KNN graphs are built for the images with high uncertainty to construct their emotion label distributions, which are based on an assumption that two adjacent facial images in the label space of the AUs recognition should have close label distributions each other. AUs recognition can depict expressions from different angles. Furthermore, AUs descriptions are less ambiguous as well as can be gained through well-developed methods effectively. Given the training set S, let $\mathbf{U}$ = $(\mathbf{u}_1^i, \mathbf{u}_2^i, \cdots, \mathbf{u}_n^i) \in R^{n\times z(0)}$ represent feature descriptions AUs, where u_n^i denotes the feature descriptions of n^{th} AU of i^{th} sample. There are numerous criteria could be utilized to measure the distance between two feature vectors, such as Eu-clidean distance, Manhattan distance, Chebyshev distance, cosine distance. The cosine distance is adopted to calculate distance of two feature vectors. Inputting an image $\mathbf{X}_i$, the approximation of the feature distance is to induce the minimization of

$$\zeta_{ij} = \begin{cases} \sum_{t=1}^{n} \frac{u_t^i \cdot u_t^j}{\|u_t^i\| \times \|u_t^j\|}, & \textit{if } \mathbf{X}_j \in N(i), \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

where $N(i)$ denotes the set of k-nearest neighbors of $\mathbf{X}_i$, and $\mathbf{X}_j$ is one of the neighbors of $\mathbf{X}_i$. The emotion label distribution is defined as

$$\mathbf{D}_i = \sum_{j=1}^{k} \frac{\zeta_{ij}}{\varpi_j} \cdot \mathbf{D}_j. \tag{10}$$

where $\varpi_j = \sum_j \zeta_{ij}$ means a normalization factor that makes sure $\sum_j d_{ij} = 1$. And $\mathbf{D}_j$ is the emotion label distribution of $\mathbf{X}_j$.

3.1.3. Label distribution learning and optimization

Given an image $\mathbf{X}_i$ with the obtained emotion label distribution $\mathbf{D}_i$, the $\mathbf{x}$=Ψ($\mathbf{X}$; $\boldsymbol{\theta}$) is introduced as the activation of the last fully connected layer in a deep CNN, where $\boldsymbol{\theta}$ denotes the vector of model parameters. Then, the softmax function is adopted to turn these activations into a probability distribution which is given as

$$p(\mathbf{y}_j|\mathbf{X}_i, \boldsymbol{\theta}) = \frac{\exp(\mathbf{x}_i)}{\sum_c \exp(\mathbf{x}_i)}. \tag{11}$$

Given the training set S= $\{(\mathbf{X}_1, \mathbf{D}_1), (\mathbf{X}_2, \mathbf{D}_2), \cdots, (\mathbf{X}_m, \mathbf{D}_m)\}$, the aim of the proposed structure is to obtain the parameter $\boldsymbol{\theta}$ to produce a distribution $\widehat{\mathbf{D}}_i$ is close to the new constructed distribution $\mathbf{D}_i$. Numerous metrics could be utilized to measure the distance between two distribution. And the Kullback–Leibler (KL) divergence is employed to quantify the distance between the new constructed emotion label distribution $\mathbf{D}_i$ and the distribution $\widehat{\mathbf{D}}_i$ from our expression recognition model, which is defined as

$$KL\left(\mathbf{D}_i \| \hat{\mathbf{D}}_i\right) = \sum_i \mathbf{D}_i \ln \frac{\mathbf{D}_i}{\hat{\mathbf{D}}_i}. \tag{12}$$

Then, the best parameter $\boldsymbol{\theta}^*$ is defined as

$$\begin{aligned} \boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_i \mathbf{D}_i \ln \tfrac{\mathbf{D}_i}{\hat{\mathbf{D}}_i} \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \sum_i \sum_j d_{ij} \ln p\left(\mathbf{y}_j | \mathbf{X}_i; \boldsymbol{\theta}\right), \end{aligned} \tag{13}$$

Thus, the loss function of our framework is determined by

$$L(\boldsymbol{\theta}) = \quad -\sum_i \sum_j d_{ij} \ln p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right). \tag{14}$$

Optimization: To obtain the minimum value of $L(\boldsymbol{\theta})$, we compute the partial derivatives of the loss function (14) adopting back propagation and update the values of each parameter using the derivatives and the learning rate. This iterative process aims to reduce the value of the loss function and make the prediction results of the model closer to the ground truth. The updating process of $\boldsymbol{\theta}_j$ can be represented as

$$\boldsymbol{\theta}_j \leftarrow \boldsymbol{\theta}_j - \alpha \frac{\delta L(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}_j}, \tag{15}$$

where α stands for the learning rate. The partial derivative of the proposed model regarding $\boldsymbol{\theta}$ could be calculated by the chain rule from the parameters of the L^{th} layer to the first layer. The recursive partial derivative equation of $\boldsymbol{\theta}$ is presented by

$$\tfrac{\delta L(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}_j} \quad = -\sum_i \sum_j d_{ij} \tfrac{1}{p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right)} \tfrac{\delta p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right)}{\delta \boldsymbol{\theta}_j}, \tag{16}$$

where

$$\frac{\delta p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right)}{\delta \boldsymbol{\theta}_j} = p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right)\left(\tau_{(j=k)} - p\left(\mathbf{y}_j | \mathbf{X}_i, \boldsymbol{\theta}\right)\right), \tag{17}$$

where $\tau(j = k)$ is 1 if $k = j$, and 0 otherwise for any k and j. Once $\boldsymbol{\theta}$ is learned, the distribution $\hat{\mathbf{D}}$ of any new sample $\mathbf{X}$ could be created by a forward propagation of the model. Eventually, the output of the proposed framework for a sample $\mathbf{X}_i$ is z^*, where

$$z^* = \underset{i}{\operatorname{argmax}} \hat{\mathbf{D}}_i. \tag{18}$$

where z^* indicates the class of facial expression output by the model. The formula symbol for the proposed method is summarized in Table 1.

4. Experiment

4.1. Datasets

Four public facial expression datasets are applied to the performance experiments of the proposed model.

RAF-DB (Li and Deng, 2018) is a large-scale and in-the-wild database that comprises 30,000 facial images collected on the Internet. The facial images are annotated with seven basic and eleven compound emotion labels by 40 trained human coders. Specifically, only 15,339 images, with the seven emotions label including neutral, anger, disgust, surprise, fear, sadness, happiness, are used in our experiment. It consists of 12,271 training images, and 3,068 testing images.

FERPlus (Barsoum et al., 2016) is also a large-scale dataset extended from FER2013, which is downloaded from the Google search engine. It consists of 28,709 training face samples, 3,589 validation facial images and 3,589 test face instances with eight emotions when the contempt is included.

Table 1
Notations used in the process of emotion label distribution construction.

Notation	Description
n	the number of *AU*
z	dimensionality of *AU* feature
c	number of categories of expression
N_i	number of *AU* appearance
$\mathbf{Y}$	probability matrix of basic expression
$\mathbf{D}_i$	label distribution matrix
$\mathbf{V}$	image level feature matrix
p_{ij}	conditional probability
L	number of layers of the model
$\sigma(\cdot)$	activation function
$\mathbf{U} \in R^{n \times z}$	feature matrix of *AU*
$\mathbf{A} \in R^{n \times n}$	correlation feature matrix
$\mathbf{H} \in R^{n \times z}$	activation matrix
$\mathbf{W} \in R^{z \times c}$	weighting matrix
$\mathbf{M} \in R^{n \times n}$	matrix of the number of occurrences of *AU*
u_i	feature representation of AU_i
y_i	probability of expression category
d_{ij}	feature representation of expression categories
g_{ij}	ground truth
f_{HYP}	hybrid pooling layer
f_{cnn}	feature vectors of the network layer
$\boldsymbol{\theta}_{cnn}$	model parameters
ε	weight parameters
ζ_{ij}	cosine distance
ϖ_j	normalization factor
S	training set
z^*	model classification result

AffectNet (Mollahosseini et al., 2017) involves over 1,000,000 human face images gathered from the Internet with 1,250 emotion-related labels, which is the largest facial expression dataset by far. Similar to FERPlus database, about 450,000 images are manually annotated with eight emotion labels. About 280,000 training samples and 4,500 test samples with seven expressions same as RAF-DB are used to evaluate the proposed model.

SFEW2.0 (Dhall et al., 2011) is built from the AFEW database by selecting key frames, which contains 879 training face samples, 406 validation instances as well as 372 testing images. And each facial image is labeled with one of the seven emotions like RAF-DB dataset.

4.2. Implementation details

4.2.1. Data pre-processing

In the label distribution construction phase, we use ResNet-50, which is pre-trained on the ImageNet dataset, as the backbone network to extract facial features. The facial features are obtained from the last pooling layer of the neural network. During the training process, we fine-tune ResNet-50 to adapt to our expression recognition task. In addition, since the methods for extracting AUs are well-developed, we employ the openface2.0 (Baltrusaitis et al., 2018) to extract AUs and their histogram of oriented gradient features in our experiments, which are then applied to build AUs graph as well as KNN graphs.

In the training phase, the MTCNN algorithm is utilized to detect facial landmarks adopted to align facial region for each facial image. Then all of them are cropped and normalized to 224 × 224 pixels. The pre-trained network is the same as the previous stage used for initializing the train model.

4.2.2. Hyperparameters selection

In the emotion label distribution construction phase, the margin δ_1 and margin δ_2 are parameters used to divide high and low uncertainty groups. The margin δ_1 can be either intended as a learnable parameter or fixed as 0.20 by default. Similar to the margin δ_1, the margin δ_2 can be either fixed as 0.32 or intended as a learnable parameter. Moreover, the ε is utilized to balance the constructed emotion label distributions with the given labels. And the ε is set to 0.80 by default. Furthermore, the λ set at 0.0001 by default based on the settings of a large number of researches.

4.2.3. Experimental settings

The proposed model is implemented by using Pytorch which is an open source machine learning library based on Torch. The experiments are performed on a PC server with two Tesla V100-SXM2-32 GB GPUs and 48 Intel(R) Xeon(R) Gold 6126 CPU@2.60 GHz. Moreover, the involved models are trained for 300 epochs. Compared to stochastic gradient descent, we consider that the momentum method has a faster convergence rate, could avoid oscillations and can jump out of the local optimal point. Therefore, Momentum-based method is adopted as the optimizer with a batch size of 128 and dropout rate of 0.5. And the momentum is set to 0.9, the weight decay is 5e-4.

4.3. Experimental results analysis

4.3.1. Results on the RAF-DB dataset

The proposed GCANet was evaluated against five popular approaches on the RAF-DB dataset, and the results are presented in Table 2. The GCANet achieved a remarkable accuracy of 88.71%, which outperforms existing methods. This success is attributed to the model's effective representation of facial expressions using geometric cues, and its use of GCNs to construct an emotion label distribution, which facilitates learning of AU expressions in a CNN framework. Additionally, the model demonstrates improved performance on uncertain samples, enabling efficient learning and improving overall accuracy. These results demonstrate the effectiveness of the proposed method in addressing the challenges of facial expression recognition in field datasets. In comparison to ReCNN, which proposes a relationship-aware facial expression recognition method based on key regions such as the eye and mouth, our method explores the mapping relationship between AUs and facial expressions at a deeper level, leveraging geometric cues between AUs to achieve superior results. Moreover, GCANet-R$_i$ (The samples only satisfy rule (i) are considered to be highly uncertain) and GCANet-R$_{ii}$ (The samples only satisfy rule (ii) are considered to be highly uncertain) obtain a lower average accuracy than GCANet (The samples that satisfy rule (i) or rule (ii) are considered to be highly uncertain). This fact proves the effectiveness of the proposed method for uncertainty suppression through constructing emotion label distribution for the samples with high uncertainty. And the performance of GCANet-R$_{ii}$ is a little better than GCANet-R$_i$. On the one hand, the some of the samples which satisfy rule (i) may meet rule (ii). On the other hand, the samples which satisfy rule (i) are less than the samples which satisfy rule (ii). In other word, although those ambiguous samples are classified correctly, they deserve more attention. Training with those samples of FER might lead to over-fitting on them. Moreover, it has a poor performance for a model to learn discriminative facial expression features.

Table 2
Comparison results expression recognition of accuray (%) with existing methods on RAF-DB database.

Methods	Classes	Validation	ACC(%)
DLP-CNN(Li and Deng, 2018)	7	10	84.22
IPA2LT(Zeng et al., 2018)	7	10	86.77
gaCNN(Li et al., 2018)	7	10	85.07
RAN(Wang et al., 2020a)	7	10	86.90
ReCNN (Xia et al., 2021)	7	10	87.06
SCN(Wang et al., 2020b)	7	10	88.14
GCANet-R$_i$	7	10	86.57
GCANet-R$_{ii}$	7	10	87.48
GCANet	7	10	**88.71**

The confusion matrices of the proposed GCANet are presented in Fig. 5. Specifically, Fig. 5(d) displays the confusion matrix on the RAF-DB dataset. The proposed GCANet approach achieves 96% accuracy for happiness expression, while attaining over 91% accuracy for sadness, surprise expressions, and neutral. Anger and disgust, on the other hand, exhibit relatively low recognition accuracies, both of which are above 82%. Fear expression exhibits the poorest performance at 79%, and is often misclassified as happiness and neutral expressions. This may be attributed to the fact that fear expression shares many similar appearance characteristics with happiness and neutral expressions, resulting in misclassifications.

4.3.2. Results on the AffectNet database

The experimental performances of the proposed method and the results comparisons with the existing methods on AffectNet are presented the database in Table 3. It can be observed that the proposed method gains a competitive accuracy that is 60.13%. Our proposed two-stage expression recognition method demonstrates excellent generalization ability on the large-scale AffectNet dataset, achieving a competitive accuracy of 60.13%. By including GCN in the emotion label construction stage, the model's ability to learn the mapping relationship between AUs and expressions during training is improved. This is attributed to the fact that the distribution of expression labels constructed from geometric facial cues is more consistent with the real label distribution of AffectNet. As a result, the model can extract deeper features related to facial expressions and improve performance. In contrast to methods such as IDFL and Tri21, which use loss functions to enhance inter-class separability and intra-class compactness, our approach focuses on accurately exploring and obtaining AU representations for face expression recognition, increasing inter-class distance and improving accuracy. SCN (Wang et al., 2020b) method achieves the highest accuracy at 60.23%, which can be attributed to its self-attention mechanism and careful relabeling mechanism. These are the only two approaches that exceed 60% accuracy. The AffectNet dataset contains various challenges such as head pose changes, occlusions, significant facial rotations, and complex environmental factors. Thus, it is hard to obtain robust features for AUs interpretation because subtle errors or inaccuracies caused by complicated factors from each procedure are accumulated. Similar to RAF-DB database, GCANet-R$_i$ and GCANet-R$_{ii}$ gain lower average accuracy than GCANet, and the performance of GCANet-R$_{ii}$ is a little higher than GCANet-R$_i$. In general, the proposed model still achieves competitive FER performance with the help of accurately analyzing geometry cues and label distribution learning.

The confusion matrix of the proposed GCANet approach on AffectNet dataset is exhibited in Fig. 5(b). The accuracy of the proposed GCANet approach achieves 78% at sadness, followed over 60% for surprise, anger and happiness. Except for neutral and contempt expression, the accuracies of remain expressions are less than 50%. Disgust and neutral expressions are the most difficult to recognize, particularly neutral expression takes the poorest performance 45%. Generally, although not getting the best results, the proposed GCANet method still improves the performance of each type of expression as far as possible.

4.3.3. Results on the FERPLUS database

The performances of our proposed GCANet on FERplus dataset compared to some of the latest methods are reported in Table 4. The experimental results demonstrate that our proposed GCANet model achieves an expression recognition accuracy of 89.25%, which is a remarkably excellent performance. We hypothesize that

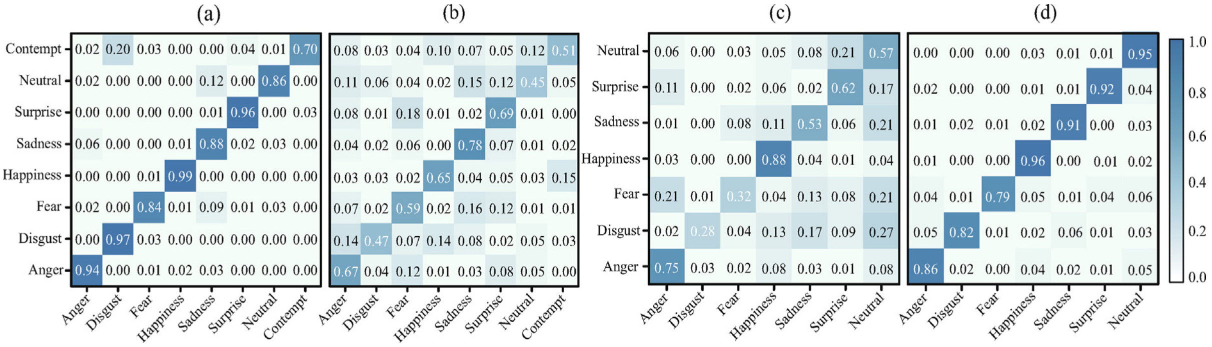

Fig. 5. Confusion matrices of the proposed GCANet on datasets. (a), (b), (c) and (d) represent FERplus, AffectNet, SFEW2.0 and RAF-DB, respectively.

Table 3
Comparison results expression recognition of accuray (%) with existing methods on AffectNet database.

Methods	Classes	Validation	ACC(%)
Upsample(Mollahosseini et al., 2017)	8	10	47.00
Weighted loss	8	10	58.00
IPA2LT‡(Zeng et al., 2018)	7	10	55.11
RAN(Wang et al., 2020a)	8	10	52.97
RAN^{+}(Wang et al., 2020a)	8	10	59.50
IDFL(Li et al., 2021)	8	10	59.20
Tri21(Xie et al., 2021)	8	10	60.12
SCN(Wang et al., 2020b)	8	10	**60.23**
GCANet-R$_{i}$	8	10	58.62
GCANet-R$_{ii}$	8	10	57.48
GCANet	8	10	60.13

Table 4
Comparison results expression recognition of accuray (%) with existing methods on FERplus database.

Methods	Classes	Validation	ACC(%)
PLD∗(Barsoum et al., 2016)	8	10	85.10
ResNet + VGG(Huang, 2017)	8	10	87.40
SeNet50∗(Albanie et al., 2018)	8	10	88.80
RAN(Wang et al., 2020a)	8	10	88.55
RAN-VGG16 (Wang et al., 2020a)	8	10	**89.55**
SCN(Wang et al., 2020b)	8	10	89.35
GCANet-R$_{i}$	8	10	87.87
GCANet-R$_{ii}$	8	10	88.59
GCANet	8	10	89.25

the accuracy is due to the precise acquisition of the relationship features between AUs and expressions in the emotion label distribution construction module. Furthermore, we obtain our newly constructed labels by applying weighted sums to the original labels, which results in labels that are closer to the true labels, thus improving the performance of the model. The highest accuracy at 89.55% is still attained by region attention network (RAN)-VGG16 (Wang et al., 2020a) model. This is largely due to the regional partial loss function, which can encourage a high degree of attention to the most valuable region. SCN (Wang et al., 2020b) achieves the second highest accuracy, which might be partly owing to its self-attention mechanism and careful relabeling mechanism. Furthermore, GCANet-R$_{i}$ and GCANet-R$_{ii}$ obtain lower average accuracy than GCANet, and the performance of GCANet-R$_{i}$ is slightly lower than GCANet-R$_{ii}$ similar to RAF-DB and AffectNet datasets. Overall, the proposed model still has learned discriminative information under complex rules on FERplus and obtains competitive FER performance.

Fig. 5(a) displays the confusion matrix of the proposed method on the FERplus dataset. Among all expressions, happiness, surprise, and disgust are relatively easier to recognize with accuracy over 96%. This may be attributed to the fact that the muscle movements are relatively more intense compared to other expressions. Furthermore, contempt is an expression that is extremely hard to identify, with the poorest performance of 70%. Note that a primary factor contributing to the low accuracy of contempt is its confusion with disgust. This is probably due to the fact that they share close muscle movements around the nose and mouth. In addition, there are two expressions that are often confused, namely sadness and fear, because these two expressions might share similar muscle deformation around the mouth.

4.3.4. Results on the SFEW2.0 database

The performance results of the proposed GCANet compared with prevalent approaches on SFEW2.0 dataset are provided in Table 5. To the best of our knowledge, our model achieves the most advanced results on the SFEW2.0 dataset with an accuracy of 56.43%. Unlike other datasets, the SFEW2.0 dataset contains facial images with substantial depth rotation, critical region occlusion, and diversity of head postures, which make it challenging to obtain discriminative information for expression classification. As a result, all experimental methods on this dataset suffer from reduced accuracy compared to other datasets, with most FER methods achieving performance under 60%. Our model addresses the issues of small sample size and large intra-class variation present in the dataset. Furthermore, the experimental results demonstrate that our proposed emotion label construction module effectively learns expression feature information in small samples. As for our models, GCANet gains higher average accuracy than GCANet-R$_{i}$ and GCANet-R$_{ii}$, and the performance of GCANet-R$_{ii}$ is a little bit higher than GCANet-R$_{i}$ similar to RAF-DB and AffectNet datasets. Finally, despite the challenges in SFEW2.0, our GCANet model further enhances the performance slightly through fusing geometry cues as well as appearance into graph network and constructing label distribution.

The confusion matrix of the GCANet on the SFEW2.0 database provided in Fig. 5(c) demonstrates that the proposed method achieves an accuracy of over 88% for happy expressions, followed by the recognition result for anger, which is over 75%. The accuracy for surprise is 62%, while the accuracy for the rest of the expression categories is less than 60%. Disgust and fear are relatively difficult for the network to recognize, with the lowest accuracy of 28% for the disgust expression. Disgust is often misclassified as sadness or neutral, and these three expressions are sometimes difficult to distinguish, most likely due to the least amount of facial movement. The confusion between fear, neutral, and anger is probably because they share similar muscle deformation around the mouth, eyes, and nose.

Table 5
Comparison results expression recognition of accuray (%) with existing methods on SFEW2.0 database.

Methods	Classes	Validation	ACC(%)
AUDN(Liu et al., 2013)	7	10	26.14
DLP-CNN(Kung et al., 2015)	7	10	51.05
EmotiW2015(Dhall et al., 2015)	7	10	56.19
GD-DNN(Mollahosseini et al., 2016)	7	10	47.70
SPDNET-4(Acharya et al., 2018)	7	10	58.14
WLS-RF(Dapogny et al., 2018)	7	10	37.10
DAUGN(Liu et al., 2019)	7	10	55.36
RAN(Wang et al., 2020a)	7	10	56.40
GCANet-R_i	7	10	53.92
GCANet-R_{ii}	7	10	54.29
GCANet	7	10	56.43

4.4. Ablation studies

4.4.1. Effectiveness of emotion label distribution construction module

To demonstrate the effectiveness of each module in GCANet, an ablation study was conducted to investigate the AUGCN-based label distribution construction and KNN graphs-based label distribution construction modules on RAF-DB. The experimental results are shown in Table 6. Several observations can be summarized as follows. First, a baseline CNN framework (ResNet-50) pre-trained on AffectNet was adopted in the first experiment (1st row). The addition of the AUGCN-based label distribution construction module (2nd row) into the basic architecture (1st row) resulted in a significant improvement in accuracy, increasing the baseline CNN framework from 84.20% to 87.96% on RAF-DB. This confirms that the AUGCN-based label distribution construction module is the most effective module for the proposed method. Moreover, the result proves that it is reasonable to learn geometry cues through the GCN architecture that was constructed to a certain extent. Furthermore, the KNN graphs-based label distribution construction module can further improve the accuracy by 0.75%. This demonstrates the necessity of designing this module and the effectiveness of uncertainty suppression.

Table 6
Evaluation of the emotion label distribution construction module of GCANet on RAF-DB database.

AUGCN-based	KNN graphs-based	Accuracy (%)
×	×	84.20
√	×	87.96
√	√	**88.71**

4.4.2. Evaluation of the ratio δ_1, δ_2 and ε

To investigate the effectiveness of various parameter configurations δ_1, δ_2 and ε utilized in the proposed method, three experiments are performed on the FER task. And the accuracy rates of these experiments on RAF-DB dataset are demonstrated in Fig. 6. The δ_2 and the ε are fixed as 0.33 and 0.80, as well as varied δ_1 in the set {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6} to train different models in the first experiment (a). The average recognition accuracies are given in Fig. 6(a). δ_1 = 0 means a sample would be assigned to the highly uncertain group if maximum prediction probability is greater than the probability of the given label. Large δ_1 results in few samples with high uncertainty. Meanwhile, small δ_1 could cause plenty of facial images with high uncertainty which may damage the performance of the model. We obtain the best accuracy rate in 0.2. In the second experiment (b), the δ_1 is fixed as 0.2 and the ε is fixed to 0.80, and evaluate δ_2 from 0.30 to 0.36. The results are illustrated in Fig. 6(b), δ_2 = 0 means that the sample is belong to the group with high uncertainty if the all prediction probabilities are equal. This suggests that the model performs the expression recognition by random guessing without learning information. Small δ_2 results in many confusing samples being divided into lowly uncertain group, which is prone to overfitting on the highly uncertain samples which might be mislabeled. Large δ_1 would lead to the easily identifiable or lowly uncertain samples being wrongly grouped into the highly uncertain group, which may greatly reduce the number of useful samples and do great harm to the model to learn useful facial expression features. The best performance is attained when δ_2 = 0.33, which confirms that the parameter should be an appropriate value. In the third experiment (c), the δ_1 is fixed as 0.2 and the δ_2 is set to 0.33, and the ε is changed in {0.5, 0.6, 0.7, 0.8, 0.9, 1.0} to learn different models. Fig. 6(c) plots the precision rates. And the new emotion label distribution constructed in the proposed GCANet is made up of two parts, i.e, emotion probability distribution and ground-truth label distribution. The parameter ε can be considered a confidence level to balance the two parts, which has a considerable impact on performance. And the best accuracy is archived when ε is set to 0.8. This indicates that the emotion probability distribution generated by our model is with high confidence, the model we utilized to construct the emotion probability distribution is reasonable.

4.4.3. Visualization analysis

To investigate the efficiency of the proposed model, the resulting 2-dimensional feature representation distributions learned from different model are visualized in Fig. 7, where the samples from the RAF-DB with various intensity are attached in different expression categories. And the t-SNE (Maaten and Hinton, 2008) is commonly used as a superior tool which is suitable for high dimensional data reduction to two or three dimensions for

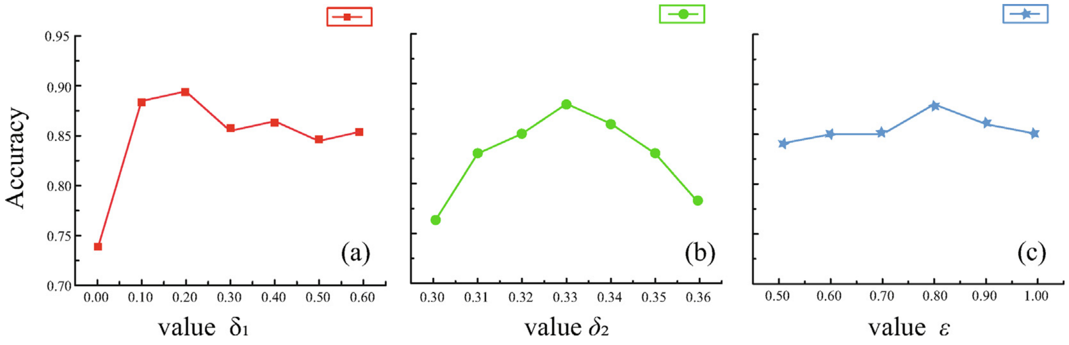

Fig. 6. Facial expression recognition accuracies with different parameters that are the margin δ_1 and δ_2, and the ratio ε on the RAF-DB dataset. (a), (b) and (c) represent δ_1, δ_2 and the ratio ε, respectively.

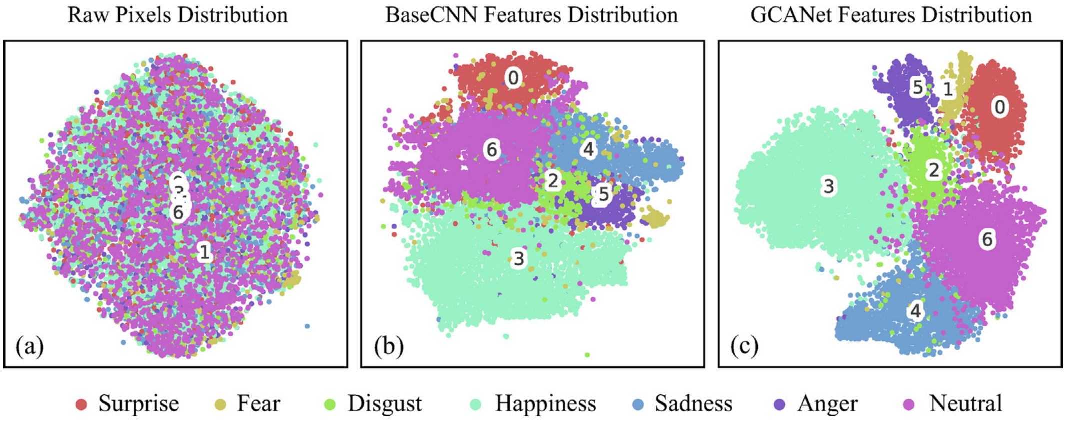

Fig. 7. Visualization study of the distribution of samples represented by (a) original pixels, (b) baseline CNN features and (c) GCANet features on the on the RAF-DB dataset.

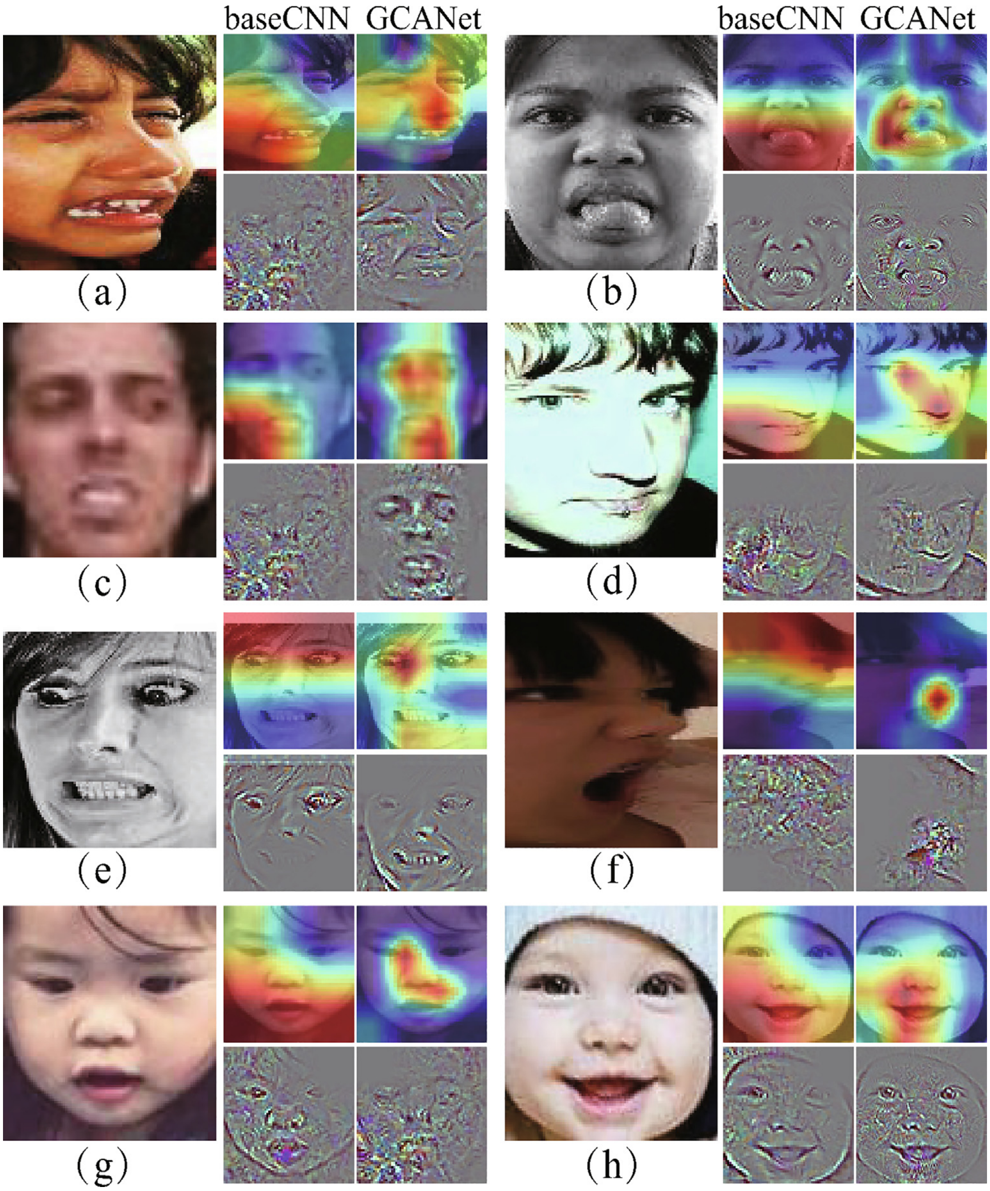

Fig. 8. Comparison of the class-discriminative regions of the different expressions on RAF-DB database. (a) Sadness. (b) Disgust. (c) Disgust. (d) Neutral. (e) Fear. (f) Anger. (g) Surprise. (h) Happiness.

visualization. The purpose of visualization comparison is to demonstrate the effectiveness of the proposed GCANet in suppressing uncertainty. In order to better describe it, Fig. 7 illustrates the distribution of facial images which are symbolized by raw pixels (original images), the feature distribution gained by baseCNN (ResNet-50) while the feature distribution obtained by the GCANet, respectively. As can be viewed, the instances which are symbolized by raw pixels are extremely irregularly distributed in the space in Fig. 7(a). The samples from RAF-DB database include various identities, poses, and lighting, and those deviations might not be restrained in original facial images. Thus, there is an enormous challenge in separating expressions of various categories in the original pixel space. In Fig. 7(b), the distribution of features obtained from the baseCNN model tends to form multiple clusters in space. Most of the samples of the same expression type appear in the same cluster. However, there are still overlapping representations in different clusters, as baseCNN may not be able to effectively suppress uncertainty due to the ambiguity of facial expressions, the subjectivity of annotators, and low-quality facial images. In contrast to the raw pixels and baseCNN feature distribution, the GCANet feature distribution in Fig. 7(c) is more clearly separated into several expression clusters. This is because the proposed method aims to preserve the local structure of the deep features, while implicitly focusing on the natural clusters in the data, and maintaining smooth transitions within clusters by accurately analyzing the potential geometric cues. This result indicates that the proposed model is effective in suppressing uncertainty.

Furthermore, Fig. 8 presents the visualization results of guided Grad-CAM maps and Grad-CAM maps (Selvaraju et al., 2017), which visualize the attentive regions of the proposed GCANet and baseCNN for some sample images. By comparing the results, it can be observed that the baseline method is distracted by irrelevant features such as hairstyles, facial outlines, and backgrounds, indicating that it is not effectively learning weight information from the discriminative regions of the face for different expressions. In contrast, the proposed GCANet is able to learn from more informative regions such as the eye, nose, eyebrow, and mouth, which contributes to its improved performance. In addition, the proposed method pays attention to the important regions while suppressing the less important regions, which facilitates more accurate learning of facial features.

5. Conclusion

This study is based on the diverse combinations of AUs for the same expression and the existence of the same AUs in different expressions. A novel and efficient EFR framework based on geometry cues-aware is proposed to suppress the uncertainty due to the ambiguity of facial expressions, the subjectivity of annotators, and low-quality facial images. This framework contains two important stages: emotion label distribution construction and learning. First, the given samples are divided into two sets (i.e., low and high uncertainty) if they meet two rules. The GCN framework is employed for the low uncertainty group to learn AUs graph representation and predict latent motion label probabilities. The emotion label distributions are subsequently produced through the blending of the given label and the prediction latent label probabilities. The k-nearest neighbor graphs are built to construct their emotion label distributions. The main idea is to identify a facial image in the low uncertainty group with the highest similarity and then replace this image. The second stage, namely the constructed emotion label distribution, would be trained directly by a CNN backbone to recognize expressions. In addition, the experimental results on four public databases, including RAF-DB, FERPlus, AffectNet, and SFEW2.0, demonstrate that the proposed method is advantageous and effective in facial expression recognition tasks.

The next step in the future work is to achieve a complete and further robust automatic facial expression recognition model with an accurate AU detection function to handle real-world uncertainty effectively.

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.

Acknowledgements

This work is supported by Major Project of the National Social Science Fund (No. 21&ZD334) and the Humanity and Social Science Youth Foundation of Ministry of Education of China (Nos. 22YJC890005).

References

Acharya, D., Huang, Z., Pani Paudel, D., Van Gool, L., 2018. Covariance pooling for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 367–374.

Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A., 2018. Emotion recognition in speech using cross-modal transfer in the wild. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 292–301.

Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P., 2018. Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp. 59–66.

Bargal, S.A., Barsoum, E., Ferrer, C.C., Zhang, C., 2016. Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 433–436.

Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z., 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D.J. et al., 2016. Interaction networks for learning about objects, relations and physics. In: Advances in Neural Information Processing Systems, pp. 4502–4510.

Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. et al., 2018. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

Benford, S., Sundnes Løvlie, A., Ryding, K., Rajkowska, P., Bodiaj, E., Paris Darzentas, D., Cameron, H., Spence, J., Egede, J., Spanjevic, B., 2022. Sensitive pictures: Emotional interpretation in the museum. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–16.

Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y., 2020. 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, pp. 13984–13993.

Chen, Y.-C., Chen, C., Martínez, R.M., Etnier, J.L., Cheng, Y., 2019. Habitual physical activity mediates the acute exercise-induced modulation of anxiety-related amygdala functional connectivity. Sci. Rep. 9, 1–12.

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, 255–271.

Dhall, A., Goecke, R., Lucey, S., Gedeon, T., 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, pp. 2106–2112.

Dhall, A., Goecke, R., Lucey, S., Gedeon, T., 2012. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia, pp. 34–41.

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T., 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426.

Donadio, M.G., Principi, F., Ferracani, A., Bertini, M., Del Bimbo, A., 2022. Engaging museum visitors with gamification of body and facial expressions. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 7000–7002.

Dong, L., Pang, Y., Song, Y., Shen, B., Bo, J., 2021a. Effects of fms program on motor skill and socialization in children with autism. In: 2021 SHAPE America Virtual National Convention & Expo. SHAPEAMERICA.

Dong, L., Shen, B., Pang, Y., Zhang, M., Xiang, Y., Xing, Y., Wright, M., Li, D., Bo, J., 2021b. Fms effects of a motor program for children with autism spectrum disorders. Percept. Mot. Skills 128, 1421–1442.

Ekman, P., 1993. Facial expression and emotion. Am. Psychol. 48, 384.

Eleftheriadis, S., Rudovic, O., Pantic, M., 2014a. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. IEEE Trans. Image Process. 24, 189–204.

Eleftheriadis, S., Rudovic, O., Pantic, M., 2014b. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. IEEE Trans. Image Process. 24, 189–204.

Fan, Y., Lam, J.C., Li, V.O.K., 2020. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In: AAAI, pp. 12701–12708.

Friesen, E., Ekman, P., 1978. Facial action coding system: a technique for the measurement of facial movement. Palo Alto 3.

Gu, J., Yang, X., De Mello, S., Kautz, J., 2017. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1548–1557.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Huang, C., 2017. Combining convolutional neural networks for emotion recognition. In: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE, pp. 1–4.

Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C. et al., 2013. Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 543–550.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kung, H.-W., Tu, Y.-H., Hsu, C.-T., 2015. Dual subspace nonnegative graph embedding for identity-independent expression recognition. IEEE Trans. Inf. Forensics Secur. 10, 626–639.

Kuruvayil, S., Palaniswamy, S., 2022. Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. J. King Saud Univ.- Comput. Informat. Sci. 34, 7271–7282.

Li, D., Wen, G., Li, X., Cai, X., 2019. Graph-based dynamic ensemble pruning for facial expression recognition. Appl. Intell. 49, 3188–3206.

Li, S., Deng, W., 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. 28, 356–370.

Li, S., Deng, W., 2019. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. Int. J. Comput. Vision 127, 884–906.

Li, Y., Lu, Y., Chen, B., Zhang, Z., Li, J., Lu, G., Zhang, D., 2021. Learning informative and discriminative features for facial expression recognition in the wild. IEEE Trans. Circuits Syst. Video Technol. 32, 3178–3189.

Li, Y., Zeng, J., Shan, S., Chen, X., 2018. Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans. Image Process. 28, 2439–2450.

Liang, X., Lin, L., Shen, X., Feng, J., Yan, S., Xing, E.P., 2017. Interpretable structure-evolving lstm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019.

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). IEEE, pp. 1–6.

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

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. Develop. Syst.

Luo, Z., Liu, L., Yin, J., Li, Y., Wu, Z., 2017. Deep learning of graphs with ngram convolutional neural networks. IEEE Trans. Knowl. Data Eng. 29, 2125–2139.

Ma, T., Tian, W., Xie, Y., 2022. Multi-level knowledge distillation for low-resolution object detection and facial expression recognition. Knowl.-Based Syst. 240, 108136.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. J. Machine Learn. Res. 9, 2579–2605.

Majumder, A., Behera, L., Subramanian, V.K., 2018. Automatic facial expression recognition system using deep network-based data fusion. IEEE Trans. Cybernet. 48, 103–114. https://doi.org/10.1109/TCYB.2016.2625419.

Mohseni, S., Zarei, N., Ramazani, S., 2014. Facial expression recognition using anatomy based facial graph. In: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 3715–3719.

Mollahosseini, A., Chan, D., Mahoor, M.H., 2016. Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp. 1–10.

Mollahosseini, A., Hasani, B., Mahoor, M.H., 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. 10, 18–31.

Nayak, S., Nagesh, B., Routray, A., Sarma, M., 2021. A human–computer interaction framework for emotion recognition through time-series thermal video sequences. Comput. Electr. Eng. 93, 107280.

Ondras, J., Celiktutan, O., Bremner, P., Gunes, H., 2020. Audio-driven robot upper-body motion synthesis. IEEE Trans. Cybernet. PP PP, 1–10. https://doi.org/10.1109/TCYB.2020.2966730.

Pantic, M., Rothkrantz, L.J.M., 2000. Automatic analysis of facial expressions: The state of the art. IEEE Trans. Pattern Anal. Machine Intell. 22, 1424–1445.

Revina, I.M., Emmanuel, W.S., 2021. Face expression recognition using ldn and dominant gradient local ternary pattern descriptors. J. King Saud Univ.-Comput. Informat. Sci. 33, 392–398.

Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M., 2012. Disentangling factors of variation for facial expression recognition. In: European Conference on Computer Vision. Springer, pp. 808–822.

Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J.M., Nasrollahi, K., Moeslund, T.B., Roca, F.X., 2017. Deep pain: Exploiting long short-term memory networks for facial expression classification. IEEE Trans. Cybernet. PP, 1–11. https://doi.org/10.1109/TCYB.2017.2662199.

Sangineto, E., Zen, G., Ricci, E., Sebe, N., 2014. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 357–366.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE Trans. Neural Networks 20, 61–80.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Tang, Y., 2013. Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.

Tian, Y.-I., Kanade, T., Cohn, J.F., 2001. Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Machine Intell. 23, 97–115.

Vasanthi, M., Seetharaman, K., 2022. Facial image recognition for biometric authentication systems using a combination of geometrical feature points and low-level visual features. J. King Saud Univ.- Comput. Informat. Sci. 34, 4109–4121.

Wang, K., Peng, X., et al., 2020a. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. 29, 4057–4069.

Wang, K., Peng, X. et al., 2020b. Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6897–6906.

Wang, X., Ye, Y., Gupta, A., 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6857–6866.

Wu, B.-F., Lin, C.-H., Huang, P.-W., Lin, T.-M., Chung, M.-L., 2017. A contactless sport training monitor based on facial expression and remote-ppg. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 846–851. https://doi.org/10.1109/SMC.2017.8122715.

Xia, Y., Yu, H., Wang, X., Jian, M., Wang, F.-Y., 2021. Relation-aware facial expression recognition. IEEE Trans. Cognitive Develop. Syst. 14, 1143–1154.

Xie, W., Shen, L., Duan, J., 2019. Adaptive weighting of handcrafted feature losses for facial expression recognition. IEEE Trans. Cybernet. 1–14. https://doi.org/10.1109/TCYB.2019.2925095.

Xie, W., Wu, H., Tian, Y., Bai, M., Shen, L., 2021. Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition. IEEE Trans. Circuits Syst. Video Technol. 32, 690–703.

Yang, H., Ciftci, U., Yin, L., 2018. Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2168–2177.

Yang, Y., Ge, S.S., Lee, T.H., Wang, C., 2008. Facial expression recognition and tracking for intelligent human-robot interaction. Intel. Serv. Robot. 1, 143–157.

Zeng, J., Shan, S., Chen, X., 2018. Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 222–237.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23, 1499–1503.

Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y., 2019. Spatial–temporal recurrent neural network for emotion recognition. IEEE Trans. Cybernet. 49, 839–847. https://doi.org/10.1109/TCYB.2017.2788081.

Zhang, Y., Wang, J., Zhang, X., 2021. Personalized sentiment classification of customer reviews via an interactive attributes attention model. Knowl.-Based Syst. 226, 107135.

Zhang, Z., Lai, C., Liu, H., Li, Y.-F., 2020. Infrared facial expression recognition via gaussian-based label distribution learning in the dark illumination environment for human emotion detection. Neurocomputing 409, 341–350.

Zhao, S., Liu, W., Liu, S., Ge, J., Liang, X., 2022. A hybrid-supervision learning algorithm for real-time un-completed face recognition. Comput. Electr. Eng. 101, 108090.

Zheng, W., 2014. Multi-view facial expression recognition based on group sparse reduced-rank regression. IEEE Trans. Affective Comput. 5, 71–85.

Zheng, W., Yue, M., Zhao, S., Liu, S., 2021. Attention-based spatial-temporal multi-scale network for face anti-spoofing. IEEE Trans. Biomet. Behav. Ident. Sci. 3, 296–307.

Zhou, Y., Xue, H.,& Geng, X., 2015. Emotion distribution recognition from facial expressions. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1247–1250.