Information Sciences 610 (2022) 707–724

Contents lists available at ScienceDirect

Information Sciences

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

ConGNN: Context-consistent cross-graph neural network for group emotion recognition in the wild

Yu Wang [a], Shunping Zhou [b], Yuanyuan Liu [b,*], Kunpeng Wang [a], Fang Fang [b], Haoyue Qian [a]

[a] School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan, China
[b] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China

ARTICLE INFO

Article history:
Received 24 March 2022
Received in revised form 30 July 2022
Accepted 1 August 2022
Available online 8 August 2022

Keywords:
Group emotion recognition
Context-consistent learning
Cross-graph neural network
Emotion bias
Multi-cue feature extraction

ABSTRACT

Group-level emotion recognition (GER) is challenging since it significantly relies on different individual facial expressions, complex group relationships, and contextual scene information. Due to complicated emotion interactions and emotion bias among multiple emotion cues, current techniques still fall short when it comes to detecting complex group emotion. In this study, we propose a context-consistent cross-graph neural network (ConGNN) for accurate GER in the wild. It can model multi-cue emotional relations and alleviate emotion bias among different cues, thus obtaining the robust and consistent group emotion representation. In ConGNN, we first extract the facial, local object, and global scene features to form multi-cue emotion features. Then, we develop a cross-graph neural network (C-GNN) for modeling inter- and intra-branch emotion relations, obtaining a comprehensive cross-branch emotion representation. To alleviate the effect of emotion bias during C-GNN training, we propose an emotion context-consistent learning mechanism with an emotion bias penalty to help obtain context-consistent group emotion, and then achieve robust GER. Furthermore, we create a new, more realistic benchmark, SiteGroEmo, and use it to evaluate ConGNN. Extensive experiments on two challenging GER datasets (GroupEmoW and SiteGroEmo) demonstrate that our ConGNN outperforms state-of-the-art techniques, with relative accuracy gains of 3.35% and 4.32%, respectively.

1. Introduction

Group emotion recognition (GER), a sub-challenge of Emotion Recognition in the Wild Challenge[1] (EmotiW), is a recent research direction of great interest in the field of affective computing and computer vision. Effective and robust GER has an important role in understanding human emotions and analyzing human intentions [1–3], and can be used in a variety of applications, such as human-computer interaction, behavior and event prediction, and smart city construction [4–6]. Rather than identifying the expressions of individual faces, GER mainly focuses on the emotional state of a group of people in complex scenes, aiming to classify the overall emotion of a group of people into three categories: positive, neutral or negative [2,7]. This requires a thorough comprehension not only of the individual's facial expression, but also of the image content and contextual information of the scene. Fig. 1 shows the differences and challenges between traditional facial expression recognition (FER) and

* Corresponding author.
E-mail addresses: vvy190701@cug.edu.cn (Y. Wang), zhoushunping@mapgis.com (S. Zhou), liuyy@cug.edu.cn (Y. Liu), wangkp@cug.edu.cn (K. Wang), fangfang@cug.edu.cn (F. Fang), qhy_2015@cug.edu.cn (H. Qian).

[1] https://sites.google.com/view/emotiw2018.

https://doi.org/10.1016/j.ins.2022.08.003
0020-0255/© 2022 Elsevier Inc. All rights reserved.

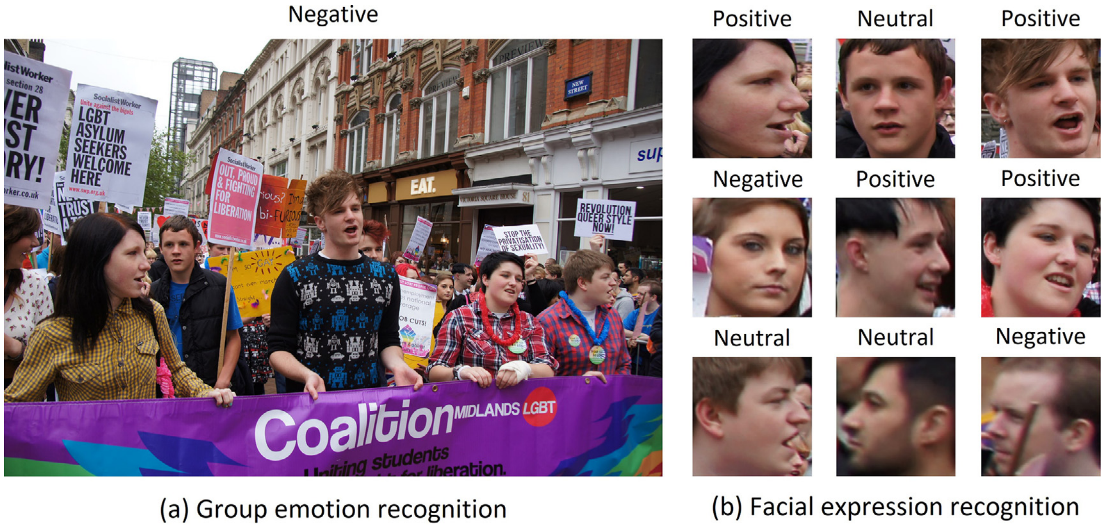


Fig. 1. Comparison of GER and FER in the wild. (a) GER of a crowd in a parade demonstration scene, and (b) FER of independent faces in the scene. Obviously, it is difficult to obtain the real group emotion in a scene by only relying on recognizing facial expressions (Fig. 1(b)), due to ignoring the situational contextual information in the scene, such as scene content, person poses, signs, etc.

GER in the wild. Compared to the traditional FER task, GER in the wild faces additional challenges, such as, undefined multiple emotion cues, complex facial expressions, crowd relations, and emotion bias among different emotion cues [8–12].

Existing methods for GER can be divided into two major categories: facial expression-based and multi-cue-based methods. Most facial expression-based methods focus on recognizing facial expressions of individuals in an image, determining the final group emotion by averaging these expressions [13]. For example, facial expression-based methods [13–15] first used the facial expression descriptor to identify the expression of each face in an image, and then employed the Group Expression Model (*GEM*) to disregard the influence of the environment and thus obtain the final group emotion. Given the considerable challenges in recognizing the overall emotion from facial expressions in a crowd, most existing facial expression-based methods only focused on the happiness expression with its corresponding intensity [11,13,14]. However, these expression-based methods are hardly adequate for the applications of GER in the wild.

The recent successful multi-cue-based methods intend to capture multiple cues, such as facial expressions and scene semantic information, in a crowd for GER [4,16,17]. Liu *et al.* [16] extracted face and overall scene context features with a deep neural network (DNN) for GER and achieved an accuracy of 71.83 % on the GAF dataset [18]. Guo *et al.* [4] extracted more emotional cues, including faces, objects, bodies, and the whole scene with a graph neural network (GNN), improving accuracy on the same dataset by 7.25 % [7]. Despite the progress achieved, these existing methods primarily focus on modeling visual information without involving effective visual relation reasoning mechanisms and emotion differences in multiple cues.

In general, GNNs [19] have been demonstrated to be particularly effective for modeling relations and importance between different cues. However, despite their potential advantages, there are two significant obstacles that make directly applying GNNs to the GER task difficult. First, different cues may have opposite emotion representations (see Fig. 2). For example, in the same image the faces show positive expressions while the scene depicts a neutral emotion. In such case, the traditional GNN relation approach cannot solve the problem of emotion bias. Second, simultaneously learning inter- and intra-branch relations is both crucial and challenging. As shown in Fig. 2(b), existing GNN methods can efficiently simulate intra-branch relations (*e.g.*, the relation between two people), but they struggle to describe the relations among different cue branches (*e.g.*, the people and the scene).

To address the aforementioned limitations and obtain context-consistent group emotions, we propose a novel context-consistent cross-graph neural network (ConGNN) for attaining robust GER in the wild. ConGNN consists of three main components, *i.e.*, the multi-branch emotion feature extractors (MFE), a cross-graph neural network (C-GNN), and emotion context-consistent learning (ECL). In particular, in MFE, we first use three parallel feature extractors, *i.e.*, facial, local object, and global scene feature extractors, to extract facial, object, and scene emotional features from different branches, respectively. Then, we employ C-GNN to model inter- and intra-branch emotion relations for a comprehensive emotion representation. To alleviate emotion bias among different branches, we introduce the ECL mechanism with an emotion bias penalty function (BPF) to make the network obtain the consistent group emotion in different emotion branches. In addition, we create a new, more realistic benchmark, and then use it to evaluate the proposed ConGNN for the GER task.

The major contributions of this study are summarized as follows.

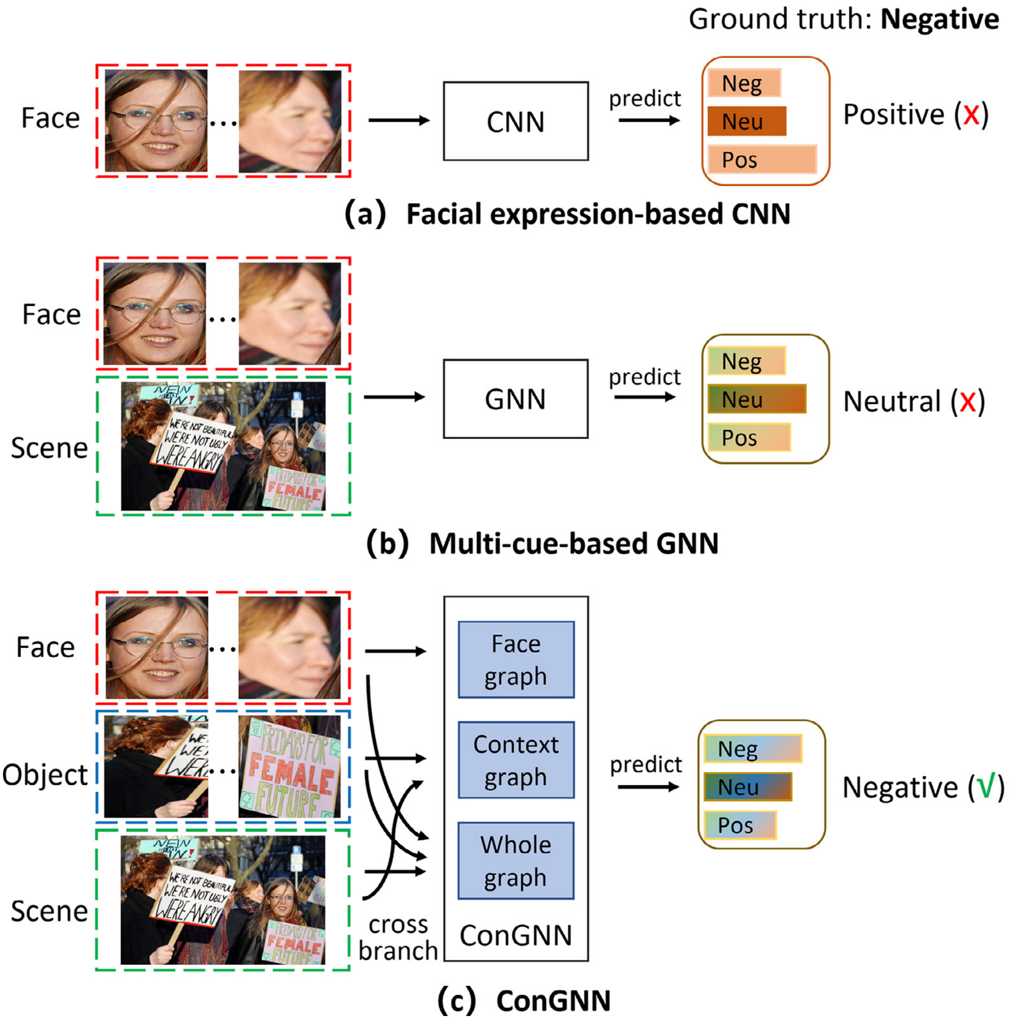


Fig. 2. Comparison of existing approaches and the proposed ConGNN for GER in the wild. (a) Facial expression-based CNN method, (b) multi-cue-based GNN method, and (c) proposed ConGNN. Inconsistent emotions in faces and scenes are a challenge for the present methodologies in (a) and (b), leading to the emotion bias and suboptimal results. The pro- posed ConGNN (see (c)) aligns the emotion bias for improving the modeling of intra- and inter-branch emotion relations, thus achieving more robust GER.

(1) We propose ConGNN to achieve accurate GER in the wild. Extensive experiments on two challenging group emotion datasets demonstrate that our approach outperforms several other widely-used techniques.
(2) To extract multiple emotion cues for comprehensive emotion representation, we introduce MFE and C-GNN for modeling the multi-cue emotion representation from the face, local object, and global scene branches. Both techniques effectively address the issues of describing inter- and intra-branch relations as well as obtaining more robust emotion representation.
(3) A novel ECL mechanism is proposed to solve emotion bias (that exists) in different branches during training. It can help the network obtain robust context-consistent group emotion learning by aligning the differences in emotion cues of various branches.
(4) We create a new, more realistic benchmark and then use it to evaluate the proposed method for the GER task. Compared with the widely used GER benchmark GroupEmoW, our benchmark has 10,034 crowd images with different countries, races, emotions, and events worldwide. It is referred to as the site group emotion (SiteGroEmo) dataset, divided into training, validation, and testing sets with 6,096, 1,972, and 1,966 images, respectively.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 presents our ConGNN approach for GER. Section 4 discusses our experimental results on publicly available and our self-built datasets. Section 5 concludes the study.

2. Related work

In this section, we first discuss the methods related to the proposed ConGNN, *i.e.*, facial expression-based, multi-cue-based, and relation learning-based methods for GER in the wild. Then, we present the datasets commonly-used in this field.

2.1. Facial expression-based methods

Facial expression-based methods recognize group-level emotion solely through the facial expressions of individuals in an image without considering the background. Given the considerable challenges in multi-person expression recognition in a crowd, early GER methods only analyzed positive emotion, *i.e.*, the happiness expression and its intensity. Herńandez *et al.* [20] calculated and averaged each individual's smile intensity in the crowd to obtain group-level happiness. Taking into account the impact of human behavior, Dhall *et al.* [21] estimated happiness intensity on the basis of the structure of group and local attributes, such as occlusion, and achieved a mean absolute error (MAE) of 0.379 on the HAPPEI dataset. Vonikakis *et al.* [15] used geometric facial features, the distribution of 100 individual expressions, and the significance of each face in a crowd to perform group-level emotion prediction. However, the aforementioned studies only considered face related information for GER, disregarding the rich scene information, which is insufficient for the effective analysis and recognition of group emotions.

2.2. Multi-cue-based methods

Recently, many studies have begun combining facial expression information with scene contextual information for GER due to the development of deep learning and group emotion datasets. In [22], group emotions were estimated by using facial expression and the whole image semantic features. Ghosh *et al.* [23] leveraged on the facial expression information, scene information, and high-level facial visual attributes for GER. More recently, Guo *et al.* [24] used the face and the whole scene features with deep CNN to perform group-level emotion prediction. Huang *et al.* [25] proposed an information aggregation method for generating feature descriptions of face, upper body, and scene for GER in the wild. Guo *et al.* [4] proposed a GNN-based model for extracting and fusing multiple emotion information, including scene, face, and object features. However, despite the positive outcomes of the multi-cue strategy, research into multi-cue extraction and fusion in the wild as well as emotion bias among multi-cues is still ongoing.

2.3. Relation learning

The relation learning framework, which has seen extensive use in computer vision and image recognition, such as picture re-ranking [26] and emotion classification [27,28], can effectively represent the relationships between objects and models [29,30]. At present, commonly-used relation learning models can be divided into two categories, namely attention-based and graph-based methods [31]. Wang *et al.* [32] proposed a cascade attention network to use the importance of each face in an image to generate a global representation for GER. Because a graph can model the relationships between nodes, GNN-based relation learning has attracted increasing interest [33,34]. Recently, more and more GNN-based methods are applied to improve the performance of GER. Guo *et al.* [4] used a GNN for understanding image emotion based on multiple cues. This GNN achieved good GER performance thanks to modeling the relationships between the emotions of the face, object, and scene. Although the aforementioned methods can help model and learn relations among multiple features, they mostly focus on intra-branch feature relations. How to adequately learn inter- and intra-branch relations remains an open research problem.

2.4. Group emotion datasets

To develop GER techniques, many of group emotion datasets in the wild have been proposed and constructed recently, *e.g.*, such as HAPPEI [13], GAF [18], GAF 2.0 [7], GAF 3.0 [2], and GroupEmoW [4]. These datasets are collected from the websites of Google, Baidu, Bing, and Flickr, crawled by some emotional keywords. Due to difficulty in labeling and acquisition, most of these datasets do not consider geographical location and scene differences. This might greatly limit the real-world application of the GER technologies. Therefore, the creation of a new GER dataset with geographical differences and information in the wild, and the development of a more robust and advantaged benchmark are highly necessary for GER tasks.

3. Methodology

In this section, we present a novel ConGNN to obtain discriminative context-consistent emotion representation, which enables robust GER in the wild [35]. Fig. 3 depicts the overall architecture of our proposed ConGNN. ConGNN consists of three main components, *i.e.*, MFE, C-GNN, and ECL. First, we use MFE to extract multi-cue emotion features from different information branches. Then, we use C-GNN for group relation learning by modeling intra- and inter- branch affective rela-

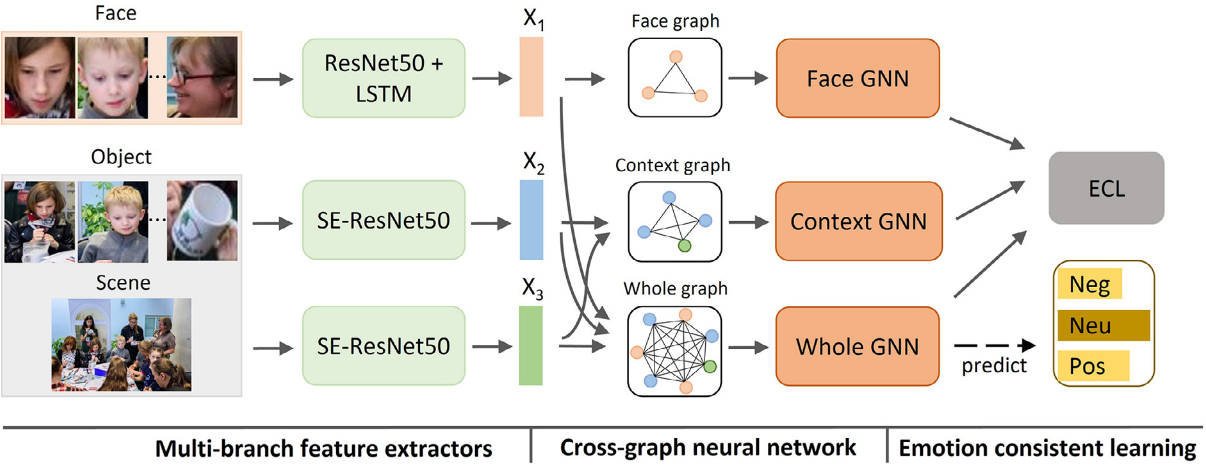


Fig. 3. Overview of ConGNN for GER in the wild. In ConGNN, we first use multi-branch feature extractors (MEF) for multi-cue emotion representation, and then use a cross-graph neural network (C-GNN) with ECL to model group relations and align emotion bias for robust group emotion prediction. Note: the dotted line only represents the prediction process.

tions. Meanwhile, a novel ECL mechanism is used to align emotion bias among different branches to help context-consistent emotion representation learning. In the succeeding sections, we subsequently explain MFE, C-GNN, and ECL.

3.1. MFE for Multi-cue emotion extraction

To obtain multi-clue emotion information from crowd scenes, we introduce three parallel feature extraction branches for respectively extracting multi-face, local object (including body and items in the scene), as well as global scene features. Three pre-trained DNNs, *i.e.*, Resnet50 [36], LSTM [37] and SE-Resnet50 [38], are employed as facial feature, object feature, and scene feature extractors.

3.1.1. Facial feature extraction

In the facial feature extraction branch, the facial regions are first detected and cropped using the standard face detectors RetinaFace [39], in order to build the face stream input. Then, we run these face regions through a pre-trained Resnet50 [36] and fine-tune them on the corresponding GroupEmoW and SiteGroEmo datasets to extract facial expression features with a size of 112 × 112. And a two-layer LSTM network is further used to learn dependence among faces. Formally, assuming that an image is p and the number of detected facial regions is N_1, we can obtain the extracted facial expression features $X_1 \in R^{L_1 \times N_1}$, which can be given by:

$$X_1 = \left[x_{11}, x_{12}, \cdots, x_{1N_1}\right], \tag{1}$$

where L_1 is the dimension of each facial expression feature.

3.1.2. Object feature extraction

For local object extraction, each image is first extracted using a bottom-up attention model, *i.e.*, the Resnet50-FPN detector [40], to acquire the salient object patches (*e.g.*, human bodies, flowers and cups) that are most related to group emotions. Then, local object features are extracted using SE-ResNet50, which is pretrained on the ImageNet-1 K database and fine-tuned on the corresponding GroupEmoW and SiteGroEmo datasets. Formally, given an image p as input, the number of detected objects is N_2 and the object emotion features $X_2 \in R^{L_2 \times N_2}$ can be written as:

$$X_2 = \left[x_{21}, x_{22}, \cdots x_{2N_2}\right], \tag{2}$$

where L_2 is the dimension of each object feature.

3.1.3. Scene feature extraction

In the global scene extraction branch, we employ the pretrained SE-ResNet50 [38] to extract the whole scene semantic features. The pretrained model is also fine-tuned on the corresponding GroupEmoW and SiteGroEmo datasets. We can obtain the extracted global scene feature $X_3 \in R^{L_3 \times 1}$, where L_3 is the dimension of the scene semantic feature.

After multi-cue feature extraction, the multi-cue emotion representation $X = \{X_1, X_2, X_3\}$ can be fed to the following C-GNN for intra- and inter-branch emotion relation modelling.

3.2. C-GNN for emotion relation learning

With the multi-cue emotion representation X, we propose C-GNN for emotion relation learning, to achieve robust comprehensive emotion representation. C-GNN is composed of two phases, *i.e.*, cross-branch graph construction and group relation learning.

3.2.1. Cross-branch graph construction

Using the multi-cue emotion features X, we initially build three complete cross-branch graphs for emotion relation learning, namely the face graph, object-context graph, and whole scene-context graph. The face graph is used to learn the relations among faces; the object-context graph is designed to establish the relations between local objects and the global scene; and the whole scene-context graph is constructed for learning the relations and interactions among all cues in the scenario, including faces, objects and the scene.

The cross-branch graph construction consists of three steps, *i.e.*, node tensor definition, message aggregation initialization, and graph construction. Fig. 4 offers a straightforward schematic view of the graph construction process.

Node tensor definition. Given each feature vector $x_{ij} \in X$ as the input, we first use a rectified linear unit (ReLU) function to normalize and project the input into an initialized node vector h^0_{ij}, and then concatenate all the node vectors of the i–th branch to form a node tensor H^0_i. They can be written as,

$$h^0_{ij} = ReLU\left(W_i x_{ij} + b_i\right), \tag{3}$$

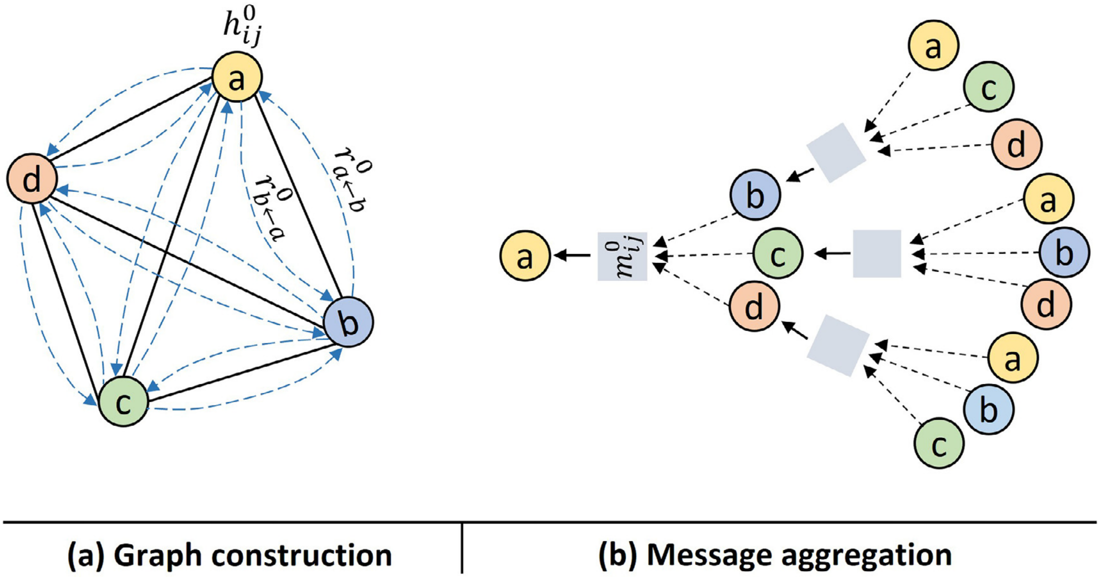
$$H^0_i = \left[h^0_{i1}, h^0_{i2}, \cdots, h^0_{iN_i}\right] \in R^{L_h \times N_i}, \quad i = 1, 2, 3, \tag{4}$$

where $h^0_{ij} \in \mathbb{R}^{L_h}$ is the j–th normalized node feature vector with the dimension of L_h. N_i is the node amount in the i–th branch. $W_i \in \mathbb{R}^{L_h \times N_i}$ and $b_i \in \mathbb{R}^{L_h}$ are the weight parameter and the bias vector of the network, respectively. Notably, W_i and b_i are shared across nodes within the same cue type.

Message aggregation initialization. As shown in Fig. 4, given the initialized node vector, we first represent the initialized message passing pairs between any node a and node b as $r(a,b) = \{r^0_{a\leftarrow b}, r^0_{b\leftarrow a}\}$, where $a, b \in \{j\}$ and $a \neq b$, which can be calculate as:

$$r^0_{a\leftarrow b} = W_b h^0_{ib}, r^0_{b\leftarrow a} = W_a h^0_{ia}, \tag{5}$$

where $W_a, W_b \in \mathbb{R}^{L_h \times L_h}$ are the weight parameter matrices associated to the corresponding nodes. h^0_{ia} and h^0_{ib} are the feature vectors of nodes a and b, respectively. Then, we aggregate all message passing to a node from its neighbor nodes as m^0_{ij}, to form the initialized message aggregation tensor $M^0_i = \left\{m^0_{ij}\right\}$, as,



(a) Graph construction

(b) Message aggregation

Fig. 4. Schematic of cross-branch graph construction. (a) The constructed graph $G = \{H^0_i, M^0_i\}$ with four nodes and six associated message passing pairs.$r^0_{a\leftarrow b}$ represents the initialized correlation message from node b to node a. (b) Message aggregation between each node and its neighbor nodes in the graph. We represent the aggregated message of the node a as $m^0_{ia} = \{r^0_{a\leftarrow b}, r^0_{a\leftarrow c}, r^0_{a\leftarrow d}\}$. Note: the message aggregation mechanism is the same in all three graph construction.

$$m_{ij}^0 = \sum_{(l)} r_{j \leftarrow l}^0, \tag{6}$$

where l represents all neighbor nodes of the node j in the graph. Fig. 4(b) shows the process of node a gathering the message vectors from all its adjacent nodes in the graph. In this manner, we obtain the final message aggregation tensor $M_i^0 = \{m_{ia}^0, m_{ib}^0, m_{ic}^0, m_{id}^0\}$ of this graph.

Cross-branch graph construction. Based on the obtained node tensor H_i^0 and message aggregation tensor M_i^0, we construct the face, object-context, and scene-context graphs in an inter-cross manner, respectively, as the following steps: 1) *for face graph construction*, we use the face node tensor H_1^0 and message aggregation tensor M_1^0 to construct the to construct the face graph $G(H_f^0, M_f^0)$ with N_1 nodes, where $H_f^0 = H_1^0$, $M_f^0 = M_1^0$; 2) *for object-context graph construction,* considering that the integration of global scene and local object features can help to suppress emotion bias among different emotional cues, we combine the node tensor of local objects H_2^0 with the node tensor of the global scene H_3^0 to construct a rich object-context node tensor $H_c^0 = \{H_2^0, H_3^0\}$. Following Eqs. (5) and (6), we can obtain the object-context message aggregation tensor M_c^0. Using H_c^0 and M_c^0, we can construct the initialized object-context graph $G(H_c^0, M_c^0)$, with $N_2 + N_3$ nodes; 3) *for whole scene-context graph construction*, we first fuse the features of all branches to form a multi-branch fused node tensor $H_w^0 = \{H_1^0, H_2^0, H_3^0\}$, and then calculate the corresponding message aggregation tensor M_w^0 via Eqs. (5) and (6), and subsequently, construct the whole graph $G(H_w^0, M_w^0)$ with $N_1 + N_2 + N_3$ nodes in the same way as above.

3.2.2. Group relation learning

Visual relationships have been proven to be crucial for many computer vision tasks [41]. To obtain the group emotion relations in varied and complex scenes, we must achieve a more comprehensive emotion representation in a large scene by interpreting and modeling relations among different emotional cues in an image. Motivated by this objective, we capture and model the intra- and inter-relations of different emotion cues in a group through C-GNN. The detailed relation learning procedure with C-GNN is provided in Algorithm 1. In C-GNN, we first employ gated recurrent units (GRUs) to iteratively

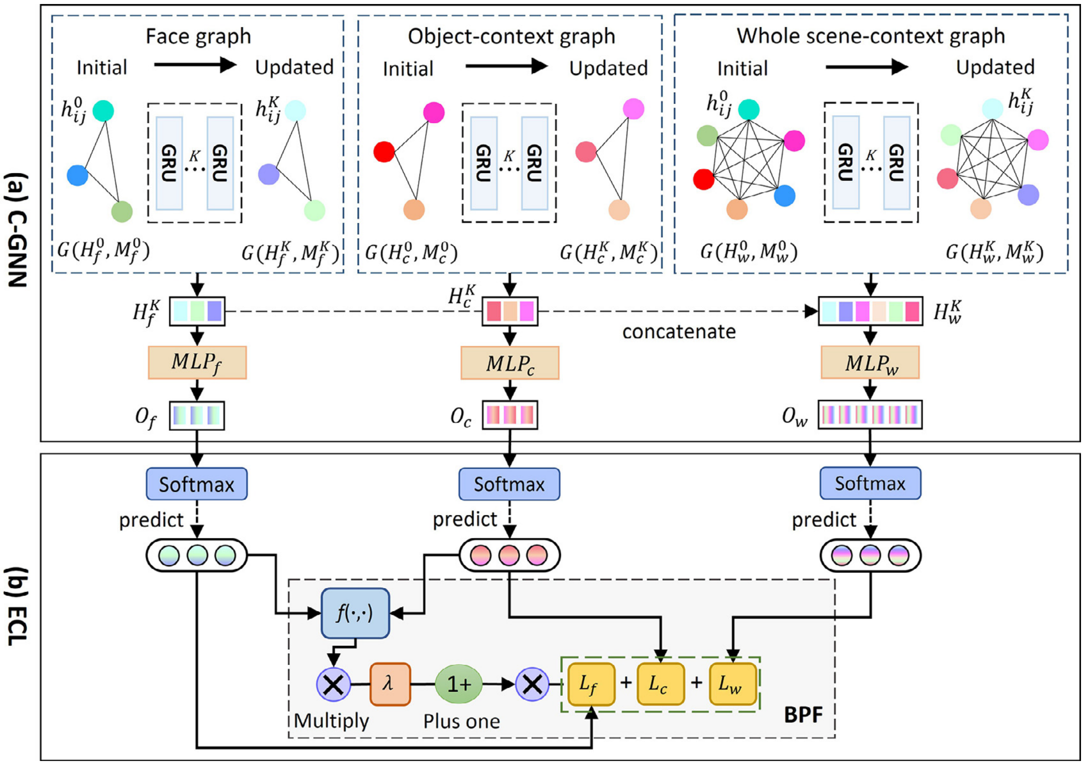


Fig. 5. Training pipeline of C-GNN with ECL. With the constructed cross-branch graphs (*i.e.*, the face graph $G(H_f^0, M_f^0)$, object-context graph $G(H_c^0, M_c^0)$, and whole scene-context graph $G(H_w^0, M_w^0)$), C-GNN first uses K-layer GRUs to model the relations of graph nodes and update each node feature in each graph of the cross-branch emotion graphs. After K iterations in GRUs, we obtain the updated graph node features H_f^K, H_c^K, H_w^K. Then, three parallel MLPs (*i.e.*, MLP_f, MLP_c, and MLP_w) are used to learn the comprehensive emotion representation $O = \{O_f, O_c, O_w\}$ from the updated node features of the graphs. Furthermore, we introduce ECL with BPF to further interact these graphs across branches, assisting C-GNN in alleviating emotion bias and achieving emotion-consistent learning. Note: h_{ij}^0 and h_{ij}^K are the initialized and updated nodes after K iterations in the graphs, respectively.

update the states of nodes and messages of the constructed graphs, and then use three multi-layer perceptions to fuse the states of graphs across branches to obtain the comprehensive emotion features $O = \{O_f, O_c, O_w\}$. The detailed training pipeline of C-GNN is depicted in Fig. 5.

More specifically, the graph updating and learning includes the following steps. Given the constructed cross-branch graphs $G(H_f^0, M_f^0)$, $G(H_c^0, M_c^0)$, and $G(H_w^0, M_w^0)$ as input, C-GNN uses the K-layer GRUs to make each node of the graph change its state all the time until the learning convergence. Through this process, message information exchanges among nodes and the relations between nodes can be modeled and learned. Fig. 6 shows the detailed architecture of GRU for graph updating in C-GNN. In the k-th iteration of GRUs, taking nodes $\{h_{ij}^{k-1}\}$ and messages $\{m_{ij}^{k-1}\}$ of previous graphs as input, the current nodes $\{h_{ij}^{k}\}$ and messages $\{h_{ij}^{k}\}$ are calculated and updated according to the procedure in Algorithm 1. Empirically, we set the number of iterative layers of GRUs to K = 4.

Algorithm 1. Detailed group relation learning procedure with C-GNN

```
**Input**:
$\{x_{ij}\}_{i=1}^{3}$: multi-cue emotion features
  $W_i, W_z, W_r, W_t, W_h, W_f, W_c, W_w$ : weight matrices
  $b_i, b_f, b_c, b_w$ : bias vectors
  $\sigma(\cdot)$: the logistic sigmoid
**Output:** comprehensive emotion features,$O = \{O_f, O_c, O_w\}$
**Initialize:** the iteration of GRUs $k \leftarrow 0$, the number of iterations: $K$
1. node feature and message aggregation vector initialization:
  $h_{ij}^0 = ReLU(W_i x_{ij} + b_i)$
  $m_{ij}^0 = \sum_{(a,b)} W_t h_{ab}^0, nodes(a,b) \neq (i,j)$
2. cross-branch graph construction:$G\left(H_f^0, M_f^0\right), G\left(H_c^0, M_c^0\right), G(H_w^0, M_w^0)$
3. graph iteration with $K$-layer GRUs
  **Repeat:**
  $k \leftarrow k + 1$
  $z_{ij}^k = \sigma(W_z \cdot \left[m_{ij}^{k-1}, h_{ij}^{k-1}\right])$
  $r_{ij}^k = \sigma(W_r \cdot \left[m_{ij}^{k-1}, h_{ij}^{k-1}\right])$
  $h_{ij}^k = \left(1 - z_{ij}^k\right) * h_{ij}^{k-1} + z_{ij}^k * tanh(W_h \cdot \left[m_{ij}^{k-1}, r_{ij}^k * h_{ij}^{k-1}\right])$
  $m_{ij}^k = \sum_{(a,b)} W_t h_{ab}^k, nodes(a,b) \neq (i,j)$
  **Until:** $k$ = $K$
4. the last graph update:$G\left(H_f^K, M_f^K\right), G\left(H_c^K, M_c^K\right), G(H_w^K, M_w^K)$
5. comprehensive emotion feature extraction:
  $O_f = W_f H_f^K + b_f; O_c = W_c H_c^K + b_c; O_w = W_w H_w^K + b_w$
```

To further model inter-branch emotion relations, at each iteration k, we first integrate the node features of the face graph and object-context graph to the whole scene-context graph via a concatenation operation, *i.e.*, $H_w^k = Concat(H_f^k, H_c^k)$. After training the GRUs with K iterations, all the nodes and their corresponding messages are updated for modelling emotional relations within branches. Then, three multi-layer perceptions denoted as MLP_f, MLP_c, and MLP_w (see Fig. 5), are employed as cross-branch emotion encoders. We represent the extracted face-branch, context-branch, and fused cross-branch emotion representations as O_f, O_c, O_w, respectively. Consequently, we have,

$$O_f = MLP_f(H_f^K), O_c = MLP_c(H_c^K), O_w = MLP_w(H_w^K), \tag{7}$$

Each of the three multi-layer perceptions MLP_f, MLP_c, and MLP_w has one fully connected layer with the LeakyRelu activation function. In this end, we obtain the comprehensive emotion features $O = \{O_f, O_c, O_w\}$. The more details can be shown in Algorithm 1.

3.3. ECL for emotion bias alignment

With the multi-cue features X and C-GNN, we can estimate the emotions of groups in the wild. However, we observe that C-GNN can focus on modeling the relations of branches and obtaining comprehensive emotion representation, while disregarding emotion bias among different branches, *e.g.*, facial expressions and scene context emotions in the same image may

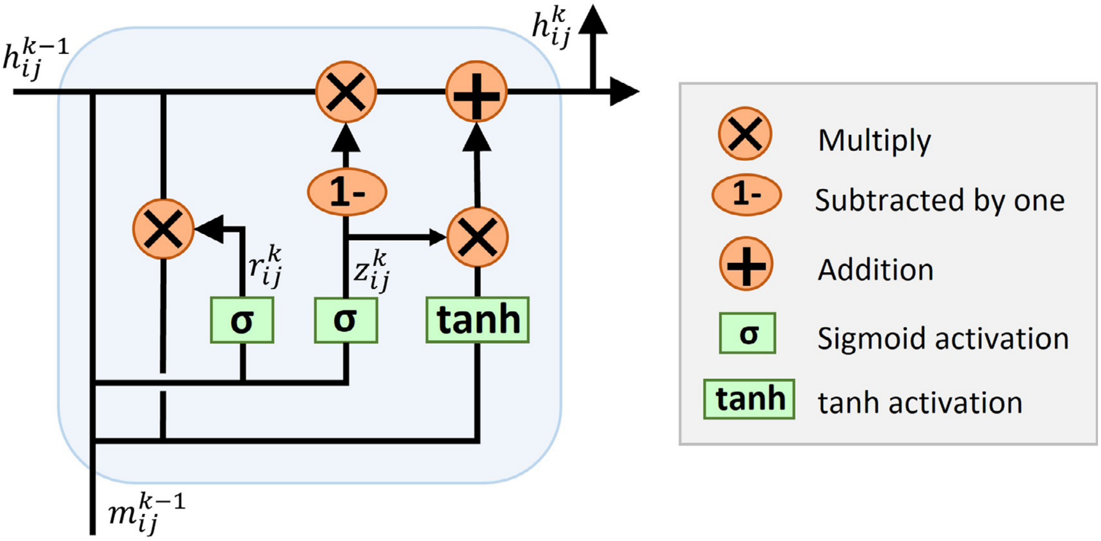


Fig. 6. The detailed architecture of the GRU. z_{ij}^{k} and r_{ij}^{k} are the reset and update gates in GRU, which control how much of the previous memory content is to be forgotten and how much of the candidate memory content is to be added.

have opposite emotion polarity. Such disregard can easily lead to incorrect emotion classification in GER. In this regard, a novel ECL mechanism and its corresponding emotion BPF are proposed to further interact these branches and help the network achieve consistent learning, and thus, alleviate the effect of emotion bias for robust GER.

As seen in Fig. 5(b), ECL with an emotion BPF includes three graph losses. The first loss is the face graph loss L_f, the second one is the object-context graph loss L_c, and the third one is the whole scene-context graph loss L_w. The cross-entropy loss is used for optimization. Mathematically, the face graph loss can be expressed as:

$$L_f = -\frac{1}{N_f}\sum_{i=1}^{N_f}\sum_{c=1}^{C}1[c = y_i]logP_{f_i,c}, \tag{8}$$

where C is the number of emotion classes (C = 3 in this study), and N_f is the number of faces. $1[c = y_i]$ is a binary indicator, and $P_{f_i,c}$ is the predicted probability that the face graph representation belongs to the group emotion c. The context graph loss L_c is given by,

$$L_c = -\frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{c=1}^{C}1[c = y_i]logP_{c_i,c}, \tag{9}$$

where N_c is the number of nodes in the object-context graph, $P_{c_i,c}$ denotes the predicted probability that the context graph representation belongs to the emotion class c. The whole fused cross-branch graph loss L_w is given by:

$$L_w = -\frac{1}{N_w}\sum_{i=1}^{N_w}\sum_{c=1}^{C}1[c = y_i]logP_{w_i,c}, \tag{10}$$

where N_w is the number of vectors in the whole scene-context graph, $P_{w_i,c}$ is the predicted probability that the fused graph representation belongs to the emotion class c. To optimize the three aforementioned losses in a consistent direction during learning, ECL introduces an emotion BPF that constrains and forces opposite direction graph loss learning for context-consistent learning. *BPF* is provided by,

$$BPF = \left(1 + \lambda * f\left(y_i^f, y_i^c\right)\right) * (L_f + L_c + L_w), \tag{11}$$

$$f\left(y_i^f, y_i^c\right) = \begin{cases} 0, if\ y_i^f = y_i^c \\ 1, if\ y_i^f \neq y_i^c \end{cases}, \tag{12}$$

where λ is a penalty coefficient that controls the penalty degree during learning. $f(.,.)$ is the penalty indicator function that indicates whether a penalty should be added. y_i^f and y_i^c are the predicted emotion categories of the face graph and object-context graph (*i.e.*, positive, negative or neutral), respectively. BPF is an adaptive consistent learning objective that effectively constrains and guides face, object-context, and whole scene-context graph losses. In summary, training C-GNN to recognize group emotions with ECL can help ensure that the information from each graph branch is properly attended to and adequately learned, resulting in consistent and robust GER.

3.4. Prediction

For inference, we only use the whole fused cross-branch emotion feature O_w to predict the group emotion. Given that cross-fusion incorporates all emotion cues, it can be used as a comprehensive emotion representation of a group for prediction. We use Softmax operation to predict the emotional class probability as follows:

$$p_c = \frac{e^{W_c \cdot O_w + b_c}}{\sum_{c=1}^{C} e^{W_c \cdot O_w + b_c}}, \tag{13}$$

where p_c is the predicted probability for the emotion class c, and C is the number of emotion categories. W_c is the c-th row of the weight matrix W of the network, and b_c is the c-th element of the bias vector b.

4. Database collection and annotation

To evaluate the proposed ConGNN method comprehensively, extensive experiments on two challenging group emotion datasets, *i.e.*, GroupEmoW [4] and SiteGroEmo, are conducted. The SiteGroEmo is a new, more realistic bench- mark collected and labeled by the authors of this paper.

4.1. SiteGroEmo: A new GER dataset

The new-established SiteGroEmo is a group-level emotion dataset with 10,034 images in the wild, collected from different tourist attractions worldwide. This dataset contains a wealth of geographic information and variation, and can be used in several downstream tasks and real-world applications, such as GER, place emotion extraction, and travel recommendation. Each image in the dataset was labeled with one of the negative, neutral and positive emotion categories. The number of negative, neutral and positive emotion categories are 1,019, 4,355 and 4,660, respectively.

Database collection. To establish the group-level emotion dataset in the wild, we collect a large amount of user-generated images from the social networking sites, *i.e.*, Flickr and Weibo platforms. These images, which depict a variety of human emotions, are taken from tourist destinations in China, Japan, Korea, Thailand, the USA, and so on. We also develop a crawler program for collecting these high-definition images from the Internet to serve as a sample source of facial expressions in the wild. After crawling the data, we eliminate images with less than two people manually, retaining group emotion images in hundreds of attractions worldwide. Finally, we collected around 15,000 images from hundreds of travel sites, including images from various locations, social environments, and events.

Database annotation. In the SiteGroEmo dataset, each photo is labeled with negative, neutral, or positive valence state by five annotators. We develop a piece of software called Expression Label Tool (ExpreLabelTool) to assist annotators in labeling efficiently. To ensure the professionalism of the annotation, five annotators trained in emotional knowledge are selected to to annotate the collected images with the tool. If more than three of the annotators give the same emotion annotation to an image, this image with the emotion annotation will be retained. Otherwise, the image will be eliminated. In the end, the dataset contains 10,034 images. For evaluation, the SiteGroEmo dataset is divided into training, validation, and testing sets with 6,096, 1,972, and 1,966 images, respectively. Fig. 7(a) shows some examples from different travel sites in the SiteGroEmo dataset.

4.2. GroupEmoW database

GroupEmoW [4] is a public GER dataset that consists of 15,894 images. It is divided into train, validation, and test sets, each with 11,127, 3,178, and 1,589 images. These images are collected from the Google, Baidu, Bing, and Flickr websites by searching for keywords related to social events, such as funeral, birthday, protest, conference, meeting, and wedding. The collective emotions of the images are also labeled with negative, neutral, or positive valence state. Fig. 7(b) shows some examples from the GroupEmoW dataset.

5. Experiments and analysis

In this section, we provide the implementation details of the proposed Con– GNN and a comparison with state-of-the-art methods.

5.1. Experimental setting and implementation details

Our ConGNN was implemented with the Pytorch and TensorFlow libraries. A warm-up mechanism was used to reduce overfitting during training. Meanwhile, a dropout scheme was used with a ratio of 0.5 in 1 × 1 convolution layers. Addition-

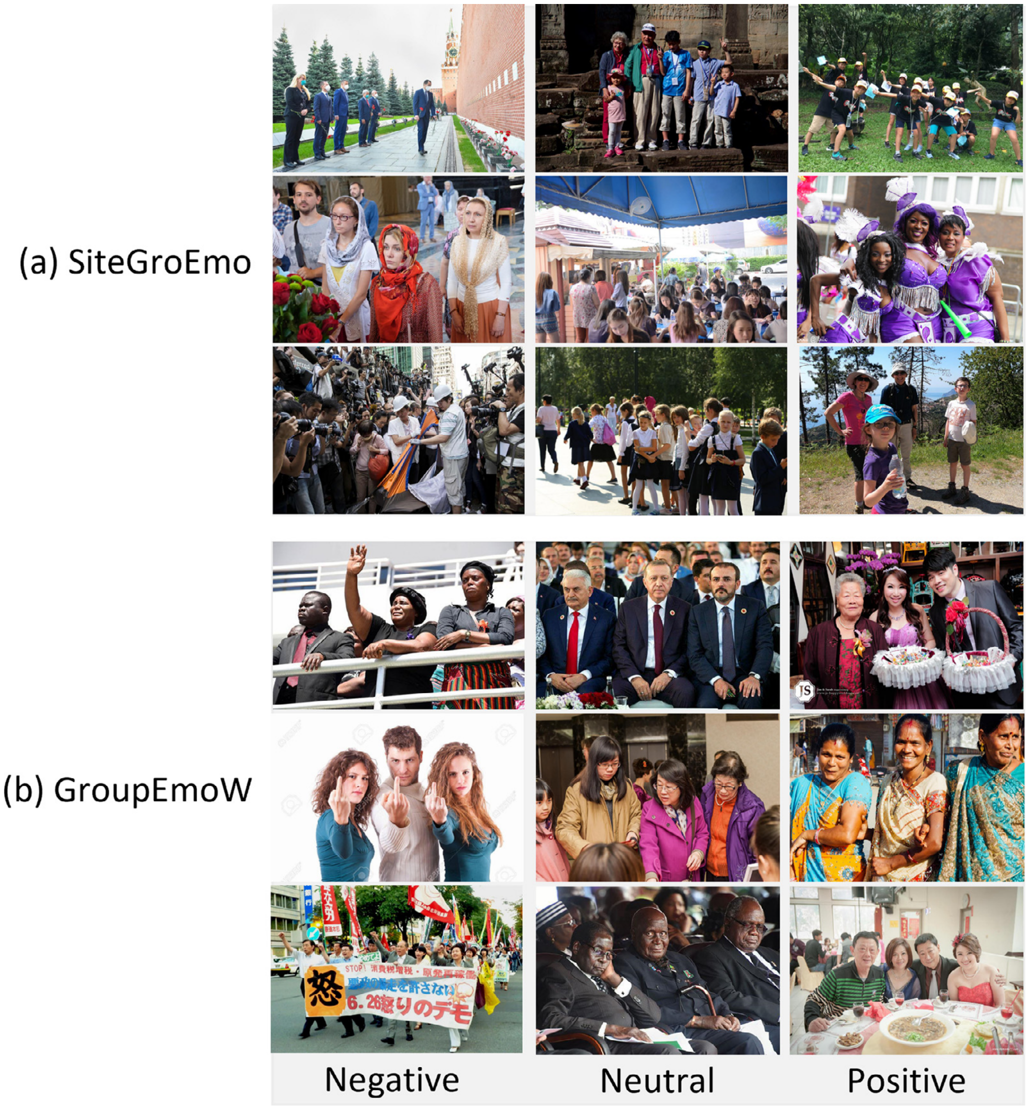


Fig. 7. Examples from (a) the SiteGroEmo dataset, (b) the GroupEmoW dataset.

ally, we employed data augmentation techniques like flips, contrast, noise, and so on. The important training parameters, *e.g.*, initial learning rate, mini-batch size and learning trick of each module are provided in Table 1.

5.2. Overall performance

5.2.1. Results on the SiteGroEmo dataset

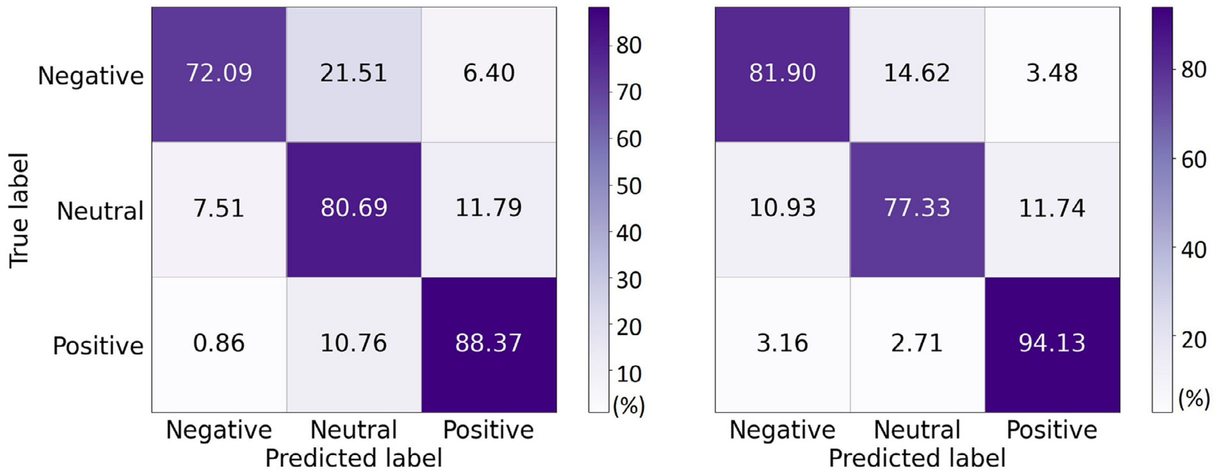
Fig. 8(a) shows the confusion matrix of ConGNN on the SiteGroEmo dataset. Among three emotion categories, the highest accuracy is 88.37 % of positive, while the lowest accuracy is 72.09 % for negative. The average accuracy of GER is 83.57 %.

To evaluate the proposed ConGNN for GER thoroughly, we compared our method with several state-of-the-art techniques, including Resnet34, SE-ResNet50 [38], Efficientnet-b2 [42], CAER-Net [43] and a GNN-based model [4]. As indicated in Table 2, our proposed method increases GER accuracy by 2.98 % and achieves the highest group emotion estimation accuracy when compared to the second-best method, *i.e.*, the GNN-based approach, which displays the best performance among other comparison methods.

In addition, to verify the robustness of our method, we conducted the 5-fold cross validation on the SiteGroEmo dataset, as shown in Table 3. Obviously, compared to the state-of-the-art methods in Table 2, our method still obtained the best performance of 82.31 %, indicating a better robustness of our method.

Table 1
Important training parameters of implementation.

Stage	Multi-branch feature extraction			Group relation learning
Network	Resnet50	LSTM	Se-Resnet50	C-GNN
learning rate	1e-4	5e-6	1e-4	2e-4
mini-batch size	32	1	32	1
learning trick	fixed learning rate			warm up



(a) SiteGroEmo (b) GroupEmoW

Fig. 8. Confusion matrices of ConGNN on (a) the SiteGroEmo, (b) the GroupEmoW datasets.

Table 2
Quantitative evaluation of ConGNN compared with other methods on the SiteGroEmo dataset. The best results are in bold.

Methods	Features	Accuracy (%)			
		Negative	Neutral	Positive	Average
Resnet34	Scene, Face	67.59	75.74	64.35	69.64
SE-ResNet50 [38]	Scene	63.82	70.53	73.99	71.58
Efficientnet-b2 [42]	Scene, Face	70.65	68.54	72.69	70.68
CAER-Net [43]	Scene, Face	**80.59**	80.02	80.46	80.28
GNN [4]	Multi-cues	73.86	**82.13**	80.41	80.59
Our ConGNN	Multi-cues	72.09	80.69	**88.37**	**83.57**

5.2.2. Results on the GroupEmoW dataset

Fig. 8(b) displays the confusion matrix of ConGNN on the GroupEmoW dataset. Among the three emotion categories, the highest accuracy is 94.13 % for positive, while the lowest accuracy is 77.33 % for Neutral. The possible reason for this result is that the neutral is difficult to distinguish from negative. The average accuracy of GER is 85.59 %.

The comparison results presented in Table 4 show that our method is still superior to other algorithms. We evaluated our method by comparing its performance with the reproduced results of Resnet34, SE-ResNet50 [38], Efficientnet-b2 [42], CAER-Net [43] and the GNN-based model [4]. As indicated in Table 4, our approach outperforms all of the other methods and has better accuracies in all three emotion categories.

5.3. Ablation study

5.3.1. Effect of different emotion cues

To evaluate the effectiveness of different emotion cues, we performed an ablation study by gradually adding different emotion cues. These inputs consisted of the extracted facial, object, and whole scene features with MFE. ConGNN was trained and tested on the GroupEmoW and SiteGroEmo datasets, respectively. The effects of different emotion cues are provided in Table 5. The best result was obtained when all the face, object, and whole scene features were used as input. Obviously, the

Table 3
5-fold cross validation on the SiteGroEmo dataset.

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Average
Our ConGNN	79.70	84.21	85.11	78.94	83.57	82.31

Table 4
Quantitative evaluation of ConGNN compared with other methods on the GroupEmoW dataset. The best results are in bold.

Methods	Features	Accuracy (%)			
		Negative	Neutral	Positive	Average
Resnet34	Scene, Face	65.17	75.74	64.35	68.13
SE-ResNet50 [38]	Scene	63.82	70.53	73.99	69.79
Efficientnet-b2 [42]	Scene, Face	71.84	72.14	73.45	72.33
CAER-Net [43]	Scene, Face	77.57	70.68	89.51	80.61
GNN [4]	Multi-cues	79.31	**80.15**	89.42	84.62
Our ConGNN	Multi-cues	**81.90**	77.33	**94.13**	**85.59**

Table 5
Ablation study on analysing ConGNN. Impacts of three different emotion cues obtained in MFE (Face, Object, Scene) on the two datasets. The best results are in bold.

Methods	Face	Object	Scene	Accuracy (%)	
				GroupEmoW	SiteGroEmo
ConGNN	√			82.82	80.11
	√	√		82.94	81.64
	√		√	84.83	81.73
	√	√	√	**85.59**	**83.57**

proposed multi-cue emotion features achieved the best performance, with an accuracy of 83.57 % on the SiteGroEmo dataset and 85.59 % on the GroupEmoW dataset, respectively.

5.3.2. Effect of different components

To better understand the role of each module in the proposed ConGNN, an ablation study was performed separately on the SiteGroEmo and GroupEmoW datasets. Table 6 presents the ablation results of the gradual addition of MFE, C-GNN and ECL components to the baseline framework. The baseline is Resnet50 that only learns the original facial features. The baseline network achieved a GER accuracy of 71.26 % and 68.74 % on the GroupEmoW and SiteGroEmo dataset, respectively. The addition of MFE can obtain an apparent improvement of 11.5 % and 12.64 %, respectively. Further integration of C-GNN improved the accuracy to 84.62 % and 82.45 %, respectively. By learning consistency among cross-branches, the addition of ECL results in relative increase of 1.15 % and 1.36 % on the GroupEmoW and SiteGroEmo datasets, respectively. With the three components, the proposed ConGNN achieved the best accuracies of 85.59 % and 83.57 % on the GroupEmoW and SiteGroEmo datasets, respectively.

5.3.3. Effect of BPF in ECL

In addition, Table 7 provides the comparison of the proposed BPF and the traditional cross-entropy (CE) loss applied to three different network models: CAER-Net [43] with facial and scene features; GNN-based method [4] with facial, object and scene features; and the proposed ConGNN with facial, object and scene features. Compared to the CE loss, the relative improvements of BPF in the three methods are 1.76 %, 0.54 % and 2.96 % on the GroupEmoW dataset, respectively, and 2.22 %, 2.19 % and 3.40 % on the SiteGroEmo dataset, respectively. The results indicate that the ConGNN framework with BPF can achieve effective context-consistent learning. We also believe that the proposed BPF can be easily extended to other machine learning applications.

5.3.4. Effect of key parameters

Effect of face amount. In this section, we discussed the effect of face amount on the MFE of the proposed method. Fig. 9 presents the accuracy of GER versus different number of faces on the SiteGroEmo dataset. For a simple and fair comparison, we only used Resnet50 [36] for evaluation. Evidently, setting the number of faces to 16 yielded the best performance. In practice, we randomly replicated the number of faces to 16 when the number of faces in an image was less than 16 and selected 16 faces with the highest confident scores when the number of faces was over 16. Additionally, we tried to use dynamic face numbers in accordance with the quantity of each input image's faces. The orange bar shows that its result is lower than the result with the number of faces set to 16.

Table 6
Ablation study of ConGNN. Effects of adding the three components (MFE, C-GNN, and ECL) to the baseline on the two datasets. The best recognition results are in bold.

Baseline	MFE	C-GNN	ECL	Accuracy (%)	
				GroupEmoW	SiteGroEmo
√				71.26	68.74
√	√			82.76	81.38
√	√	√		84.62	82.45
√	√	√	√	**85.59**	**83.57**

Table 7
Performance comparison between CE and BPF.

Method	Loss	Accuracy (%)	
		GroupEmoW	SiteGroEmo
CAER-Net [43]	CE	80.61	79.13
	BPF	82.03	82.28
GNN-based [4]	CE	84.62	81.31
	BPF	85.08	82.45
Our ConGNN	CE	83.13	80.82
	BPF	**85.59**	**83.57**

Effect of the penalty coefficient. As described in section 3.3, the penalty coefficient λ controls the penalty extent in BPF. To further assess the impact of penalty coefficient λ in the BPF, we trained ConGNN with different λ values. Fig. 10 shows the performance of ConGNN under different λ values on the GroupEmoW (orange) and SiteGroEmo (blue) datasets, respectively. It is clear that, the best accuracy can be obtained on both two datasets, when λ is set to 0.2. And with $\lambda = 0.2$, we obtained the best accuracy of 85.59 % and 83.57 % on the GroupEmoW and SiteGroEmo, respectively.

5.3.5. Cross-database experiments on GroupEmoW → SiteGroEmo

In addition, to verify the generalizability of ConGNN, cross-database validation was conducted on the challenging in-the-wild SiteGroEmo dataset. Images from the GroupEmoW dataset were used for training, whereas images from the SiteGroEmo testing set were used for testing without fine-tuning. Table 8 presents the comparison results of the proposed model and state-of-the-art methods, including ResNet34, SE-ResNet50 [38], Efficientnet-b2 [42] and GNN-based model [4]. Although the training and testing datasets have different settings (*e.g.*, scene, pose, lighting, ethnicity, age, etc.), the results of ConGNN demonstrate that it is reusable for group emotion recognition on the SiteGroEmo dataset. Our method achieved an accuracy of 76.24 %, gaining relative 30.2 % and 0.49 % improvements over the recognition accuracies of CAER-Net and GNN, respectively.

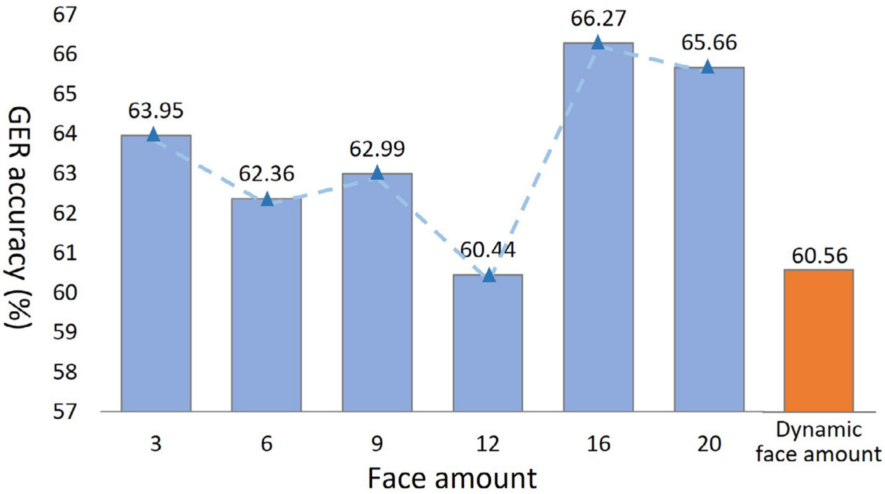


Fig. 9. Effect of different numbers of faces in an image. For a simple and fair comparison, we only use Resnet50 for evaluation.

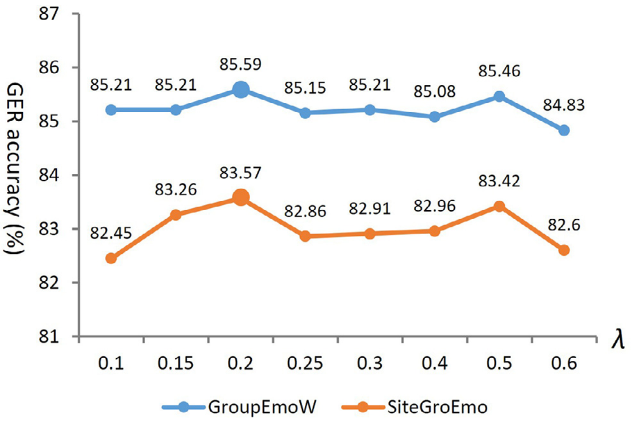


Fig. 10. Accuracy comparison of GER with different penalty coefficient λ values in BPF on the GroupEmoW (orange) and SiteGroEmo (blue) datasets, respectively.

Table 8
Cross-validation comparison with state-of-the-art methods on GroupEmoW → SiteGroEmo. The best results are in bold.

Methods	Features	Accuracy (%)
Resnet34	Scene, Face	61.06
SE-ResNet50 [38]	Scene	45.64
Efficientnet-b2 [42]	Scene, Face	56.43
CAER-Net [43]	Scene, Face	58.54
GNN [4]	Multi-cues	75.87
Our ConGNN	Multi-cues	**76.24**

5.4. Visualization

5.4.1. Visualization of different emotion representations

To evaluate the effect of different emotion representations, we visualized the cross-branch emotion representation O_w w/o the ECL and multi-cue emotion representation X in 2D feature space by using the *t*-SNE [44] on the GroupEmoW dataset.

Compared to the results in Fig. 11(a), the multi-cue emotion representation O_w achieved closer intra-class distances and longer inter-class distances (see Fig. 11(b)). With the further addition of ECL to O_w, the cross-branch emotion representation O_w obtained the best clustering, as shown in Fig. 11(c). This result suggests that both C-GNN and ECL technologies can successfully alleviate the influence of the emotion bias among multi-feature information, learning a more robust and discriminative emotion representation.

5.4.2. Visualization of emotion bias

Fig. 12 visualize the emotion bias in various models, namely, Resnet50 [36], GNN-based model [4], and our ConGNN, with different emotion cues. F represents the face cue, S represents the scene cue, and M–C represents the multi-cues proposed in

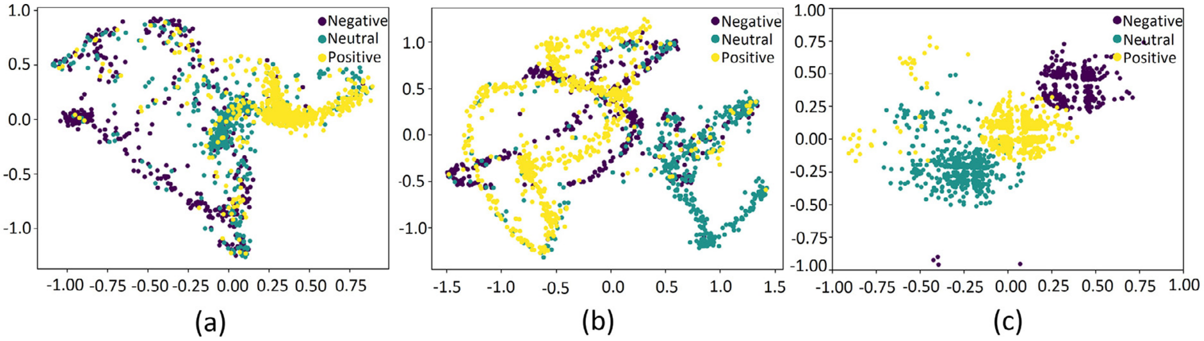


Fig. 11. 2D *t*-SNE visualization of different emotion representations. (a) Multi-cue feature X, (b) cross-branch feature O_w without using ECL, and (c) cross-branch emotion feature O_w with ECL.

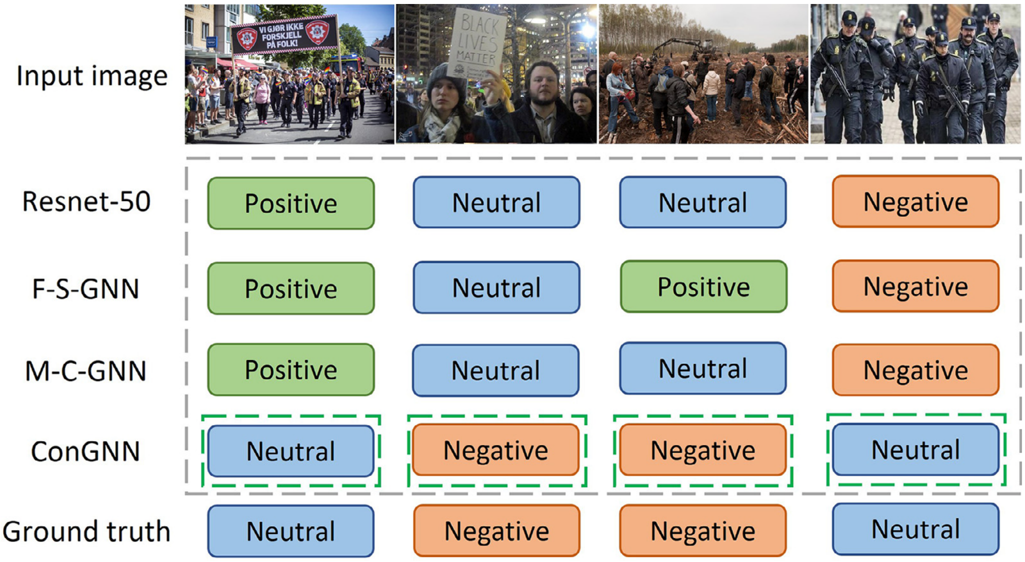


Fig. 12. Emotion bias in different models. The green dashed boxes indicate the correct results.

Table 9
Computational cost of training and testing with different methods on the two datasets. FPS is frames per second.

Methods	GroupEmoW (FPS)		SiteGroEmo (FPS)	
	Training	Testing	Training	Testing
Multi-cue-concat	44	93	58	94
GNN [4]	41	92	52	93
Our ConGNN	30	91	39	94

this study. The results demonstrate that our ConGNN with multiple emotion cues can effectively alleviate the emotion bias and obtain the best performance, compared with other models.

5.5. Time complexity

Table 9 reports the training and testing time complexities of GER by using different methods, including Multi-cue-concat, GNN-based model [4], and our ConGNN, on the GroupEmoW and SiteGroEmo datasets, respectively. Multi- cue-concat means that we just concatenate the facial, object and scene features for GER. Referred to Liu *et al.* [45], the training complexity referred to the time for one backpropagation during training. All experiments were conducted on a PC with Intel Core i7-10700 CPU at 2.90 GHz, 16 GB memory, and NVIDA GeForce GTX 2070 SUPER. Our proposed ConGNN resulted in the best performance with only a small additional computational cost (91 FPS), indicating that the proposed method exhibits improved accuracy and efficiency. Compared to the GNN, the proposed model achieved the improvement of inference speed of 1 FPS, indicating that our method does not introduce much computational cost.

6. Conclusion

This study proposed ConGNN, which can mitigate the influence of emotion bias among different emotion information (*i.e.*, individual facial expressions, object emotions, and scene emotions) for robust GER in wild scenes. We first used three feature extractors to extract multi-cue emotion features in three branches, and then proposed the novel C-GNN with ECL mechanism to learn inter- and intra-branch group affective relations and obtained a comprehensive cross-branch emotion representation. ECL can help the network alleviate the influence of emotion bias, and thus, achieve robust GER. Additionally, we built a new SiteGroEmo dataset for the evaluation of ConGNN. The extensive experiments on the GroupEmoW and SiteGroEmo datasets demonstrated that ConGNN achieved relative performance improvements of 3.35 % and 4.32 %, respectively. In the future, we will further consider geographic location and event information into our ConGNN to obtain more accurate group emotion recognition.

CRediT authorship contribution statement

Yu Wang: Data curation, Software, Visualization, Methodology, Writing – original draft. **Shunping Zhou:** Validation, Supervision, Writing – review & editing. **Yuanyuan Liu:** Conceptualization, Methodology, Validation, Project administration, Writing – review & editing. **Kunpeng Wang:** Data curation, Software. **Fang Fang:** Writing – review & editing. **Haoyue Qian:** Writing – review & editing.

Data availability

Data will be made available on request.

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 was supported by the National Natural Science Foundation of China grant (62076227, U1613216) and Wuhan Applied Fundamental Frontier Project grant (2020010601012166).

References

[1] K. Fujii, D. Sugimura, T. Hamamoto, Hierarchical group-level emotion recognition, IEEE Trans. Multimedia 23 (2021) 3892–3906.

[2] A. Dhall, A. Kaur, R. Göcke, T. Gedeon, Emotiw 2018: Audio-video, student engagement and group-level affect prediction, Proceedings of the 20th ACM International Conference on Multimodal Interaction (2018) 653-656.

[3] G.A. van Kleef, A.H. Fischer, Emotional collectives: How groups shape emotions and emotions shape groups, Cognition and Emotion 30 (1) (2016) 3-19, pMID: 26391957.

[4] X. Guo, L. Polania, B. Zhu, C. Boncelet, K. Barner, Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2921–2930.

[5] A.A. Seate, D. Mastro, Exposure to immigration in the news: The impact of group-level emotions on intergroup behavior, Communication Research 44 (6) (2017) 817–840.

[6] D. Yu, L. Xingyu, D. Shuzhan, Y. Lei, Group emotion recognition based on global and local features, IEEE Access 7 (2019) 111617–111624.

[7] A. Dhall, R. Gocke, S. Ghosh, J. Joshi, J. Hoey, T. Gedeon, From individual to group-level emotion recognition: Emotiw 5.0, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 524–528.

[8] Y. Liu, W. Dai, F. Fang, Y. Chen, R. Huang, R. Wang, B. Wan, Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition, Inf. Sci. 578 (2021) 195–213.

[9] L. Chen, M. Zhou, W. Su, M. Wu, J. She, K. Hirota, Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction, Inf. Sci. 428 (2018) 49–61.

[10] F.Z. Canal, T.R. Muller, J.C. Matias, G.G. Scotton, A.R. de Sa Junior, E. Pozzebon, A.C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, Inf. Sci. 582 (2022) 593–617.

[11] L. Evtodienko, Multimodal end-to-end group emotion recognition using cross-modal attention, arXiv preprint arXiv:2111.05890 (2021).

[12] J. Yu, M. Tan, H. Zhang, Y. Rui, D. Tao, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2) (2022) 563–578.

[13] A. Dhall, R. Goecke, T. Gedeon, Automatic group happiness intensity analysis, IEEE Trans. Affective Comput. 6 (1) (2015) 13–26.

[14] X. Huang, A. Dhall, G. Zhao, R. Goecke, M. Pietikäinen, Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis., in: BMVC, 2015, pp. 34-1.

[15] V. Vonikakis, Y. Yazici, V.D. Nguyen, S. Winkler, Group happiness assessment using geometric features and dataset balancing, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 479–486.

[16] N. Liu, Y. Fang, Y. Guo, Enhancing feature correlation for bi-modal group emotion recognition, Pacific Rim Conference on Multimedia, Springer (2018) 24–34.

[17] A.S. Khan, Z. Li, J. Cai, Y. Tong, Regional attention networks with context-aware fusion for group emotion recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1150–1159.

[18] A. Dhall, J. Joshi, K. Sikka, R. Göcke, N. Sebe, The more the merrier: Analysing the affect of a group of people in images, 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 1 (2015) 1-8.

[19] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: European Semantic Web Conference, Springer, 2018, pp. 593–607.

[20] J. Hernandez, M. Hoque, W. Drevo, R.W. Picard, Mood meter: counting smiles in the wild, in: In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 301–310.

[21] A. Dhall, J. Joshi, I. Radwan, R. Goecke, Finding happiest moments in a social context, in, Asian Conference on Computer Vision, Springer (2012) 613–626.

[22] Q. Wei, Y. Zhao, Q. Xu, L. Li, J. He, L. Yu, B. Sun, A new deep-learning framework for group emotion recognition, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 587–592.

[23] S. Ghosh, A. Dhall, N. Sebe, Automatic group affect analysis in images via visual attribute and feature networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1967–1971.

[24] X. Guo, L.F. Polanía, K.E. Barner, Group-level emotion recognition using deep models on image scene, faces, and skeletons, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 603–608.

[25] X. Huang, A. Dhall, R. Goecke, M. Pietikainen, G. Zhao, Multimodal framework for analyzing the affect of a group of people, IEEE Trans. Multimedia 20 (10) (2018) 2706–2721.

[26] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, IEEE Trans. Multimedia 16 (1) (2014) 159–168.

[27] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, Z. Luo, Clip-aware expressive feature learning for video-based facial expression recognition, Inf. Sci. 598 (2022) 182–195.

[28] M. Hou, Z. Zhang, Q. Cao, D. Zhang, G. Lu, Multi-view speech emotion recognition via collective relation construction, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 218–229.

[29] G. Hu, B. Cui, Y. He, S. Yu, Progressive relation learning for group activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 980–989.

[30] L. Cheng, F. Xie, J. Ren, Kb-qa based on multi-task learning and negative sample generation, Inf. Sci. 574 (2021) 349–362.

[31] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.

[32] K. Wang, X. Zeng, J. Yang, D. Meng, K. Zhang, X. Peng, Y. Qiao, Cascade attention networks for group emotion recognition with face, body and image cues, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 640–645.

[33] N. Chairatanakul, X. Liu, T. Murata, Pgra: Projected graph relation- feature attention network for heterogeneous information network embedding, Inf. Sci. 570 (2021) 769–794.

[34] S. Xu, Y. Xiang, Frog-gnn: Multi-perspective aggregation based graph neural network for few-shot text classification, Expert Syst. Appl. 176 (2021) 114795.

[35] J. Zhang, Y. Cao, Q. Wu, Vector of locally and adaptively aggregated descriptors for image feature representation, Pattern Recogn. 116 (2021) 107952.

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

[37] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Trans. Neural Networks Learn. Syst. 28 (10) (2017) 2222–2232.

[38] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[39] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, S. Zafeiriou, Retinaface: Single- stage dense face localisation in the wild, ArXiv abs/1905.00641 (2019).

[40] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[41] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10313–10322.

[42] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, 2019, pp. 6105–6114.

[43] J. Lee, S. Kim, S. Kim, J. Park, K. Sohn, Context-aware emotion recog- nition networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10142–10151.

[44] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (11) (2008) 2579–2605.

[45] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recogn. 84 (2018) 251–261.