A Self-Fusion Network Based on Contrastive Learning for Group Emotion Recognition

Xingzhi Wang, Dong Zhang, Hong-Zhou Tan, *Senior Member, IEEE*, and Dah-Jye Lee

***Abstract*— Group emotion recognition (GER) from image has attracted much attention in recent years. Networks using attention mechanism for GER have shown great potential. However, the performance of the current attention-based GER networks suffers from the indistinctive features of individuals in the group, poor feature fusion weights, and the lack of semantic information of the objects in the image. We present a new framework that is composed of three networks, FacesNet, SceneNet, and ObjectsNet, to address these shortcomings. This new framework is designed to recognize group emotion by exploiting the information from the faces, scene, and objects in image. In FacesNet, we use contrastive learning to help the network extract distinctive emotion features and a new attention mechanism named self-fusion module to generate precise fusion weights for aggregation of individual facial features. We design SceneNet to capture the multiscale scene features to exploit the emotion cues from the scene. We construct a fully connected network named ObjectsNet to classify the semantic features of the objects. Finally, we linearly integrate the outputs of these three networks as the final output of this unique framework for GER. Experiment results on three datasets for GER show that our proposed framework achieved better performance in terms of recognition accuracy compared with the state-of-the-art methods.**

***Index Terms*— Attention mechanism, contrastive learning, group emotion, recognition.**

I. Introduction

EMOTION recognition is the process of identifying human emotion based on facial expressions from video or image, spoken expression from audio, written expression from text, and physiology as measured by wearables. Automatic emotion recognition plays an important role in wide applications, such as detection of depression [1], abnormal event detection [2], and human–computer interaction [3]. Individual emotion recognition (IER) focuses on classifying the image [4], [5], video [6], [7], audio [8], [9], or biological signals [10] of an individual into typical emotions, e.g., anger, disgust, fear, happiness, sadness, and surprise, while group emotion recognition (GER) usually aims to classify the overall group emotion of a number of people into three classes, including positive, neutral, and negative [11]. GER is more challenging than IER as there are complex relationships involving interactions among the people in the group and the environment they are in.

A group is formed by individuals and the emotions of individuals form the state of group emotion. A critical question lying in GER research is how to aggregate the individual emotion features for GER. With the advance of machine learning especially deep learning technology, researchers in image-based GER have proposed many promising networks to recognize group emotion by fusing the prediction results or features of the facial expressions, scene, and the surrounding objects. Some researchers considered individuals in a group contribute equally to group emotion and described group emotion by averaging the individual emotion features or using a voting mechanism to yield a unique result of recognition [12], [13], [14], [15]. There are also works that fuse individual features together by long short-term memory (LSTM) [16], [17], gated recurrent units (GRUs) [18], or graph neural network (GNN) [18] for GER.

In recent years, the success of attention mechanism in computer vision made it a promising technique for many recognition tasks. Attention mechanism has also been used to perform image-based GER to generate the attention weights for individual features aggregation explicitly. The attention-based GER methods usually use the classification loss to supervise the learning of individual emotion features, the output of a sigmoid function as fusion weights to characterize the importance of individual features in fusion, and high-level feature maps of the whole image for recognition. The experimental results show that the attention-based methods achieved the state-of-the-art performance in terms of recognition accuracy [19], [20], [21].

Current attention-based GER methods still face three challenges. First, the individual emotion features are not distinctive because the learning process is only supervised by the loss calculated from the final classification result. Second, the sigmoid function used for generating fusion weights is not able to provide precise representation of the importance of individual features. As shown in Fig. 1, the sigmoid function approximates the input in two large slow changing areas to either 0 or 1 to provide a very coarse representation (weight) of the importance of individual features in the attention module of

Manuscript received 12 February 2022; revised 6 May 2022 and 27 June 2022; accepted 23 August 2022. Date of publication 12 September 2022; date of current version 3 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62173353; in part by the Guangzhou Municipal People's Livelihood Science and Technology Plan under Grant 201903010040; and in part by the Science and Technology Program of Guangzhou, China, under Grant 202007030011. *(Corresponding author: Dong Zhang.)*

Xingzhi Wang, Dong Zhang, and Hong-Zhou Tan are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: wangxzh58@mail2.sysu.edu.cn; zhangd@mail.sysu.edu.cn; issthz@mail.sysu.edu.cn).

Dah-Jye Lee is with the Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT 84602 USA (e-mail: djlee@byu.edu).

Digital Object Identifier 10.1109/TCSS.2022.3202249

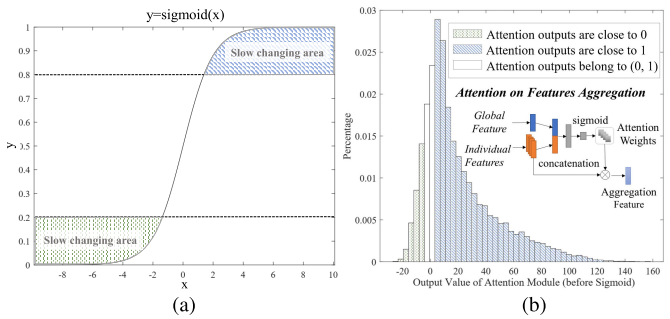


Fig. 1. (a) Sigmoid activation function graph. The sigmoid function maps only a small region ([−3, 3]) of the input x to the range of [0, 1], but approximates two large regions [(− ∞, −3] and [3, + ∞)] of input to either 0 or 1. (b) Histogram of the input to the sigmoid activation of the attention module in [19] using the test set of the GroupEmoW [18] dataset. Only a small portion of the input values are mapped to [0, 1]. Most of the distinctive input values are approximated to either 0 or 1 (slow changing areas).

the GER networks. Third, although the semantic information of the detected objects is a good indication of the group emotion and has the potential to help improve the recognition accuracy, it has not been used in related GER research. These observations affirm that the performance of the attention-based GER methods can be further improved.

In this article, we propose a contrastive learning-based self-fusion network to recognize group emotion in image. We use contrastive learning to help the network efficiently extract distinctive emotion-specific features for classification. The self-fusion module is designed to aggregate individual emotion features to generate precise weights for feature fusion. It maps the global (scene) emotion features and individual emotion features into a common nonnegative subspace and uses the cosine similarity between the global and individual emotion features in the subspace as the fusion weights. In addition, we use a bag-of-word (BoW) [22] model to code the detected semantic labels of a group and construct a fully connected network to classify group emotion. The experiments on three public benchmark datasets show that the proposed method obtained better performance in terms of group emotion recognition accuracy compared with the state-of-the-art approaches.

The remainder of this article is structured as follows: In Section II, we briefly review related works in GER. In Section III, we elaborate on the proposed method in detail. The experimental results and extensive comparisons are reported in Section IV. Finally, we summarize our work in Section V.

II. Related Work

A. Individual Emotion Recognition

IER, as the foundation of GER, has received wide attention in the past decades. In recent years, facilitated with the improvement of deep learning, the IER approaches have achieved promising performance. Compared with text [23], audio [8], [9], and biological signal [24], [25], facial expression is believed to be the most powerful information for IER and the most straightforward way for human to convey their emotion [26]. It makes facial expression recognition (FER) one of the popular research topics for automatic emotion recognition. As the foundational research of image-based GER, many promising FER methods have been presented to determine the individual emotion from a facial image.

For instance, Levi and Hassner [27] used LBP features as the input of their deep model for illumination-invariant FER. Zhang *et al.* [5] used SIFT features to promote the robustness of the learned model against image scaling and rotation. Wang *et al.* [28] proposed a multitask FER network to alleviate the impact from pose variations and identity, and used an adversarial discriminator to extract discriminative features for FER. Liu *et al.* [29] presented a boosted deep belief network (BDBN) to integrate feature representation, feature selection, and classification boosting into a loopy framework for FER. The success of these approaches manifests that facial expression provides a powerful indication of individual emotion. The technique of IER forms the foundation of GER.

B. Group Emotion Recognition

GER, which determines the emotion of a group of people, has made significant progress in recent years with the flooding of data on social media. The pioneering work [30] of GER labeled the group-level emotion as "Positive," "Negative," or "Neutral" and released the Group Affect Database, which contains images of a group of people in social events. The EmotiW group-level emotion recognition subchallenges provided a larger image dataset and dramatically promoted GER research [31], [32]. Several studies have been presented to tackle the GER task. For example, Tan *et al.* [14] assumed that the emotion of group is a simple superposition of individual emotions and yielded group-level emotion by averaging the prediction of aligned and nonaligned faces. Furthermore, Gupta *et al.* presented an attention model to generate attention weights for individuals in a group and used the weighted average of the individual facial features as a group-level feature for GER. The research works of [18] and [20] also used variants of the attention module [19] to improve the recognition accuracy of their proposed models. Recently, more aggregation techniques, such as LSTM and GNN, were used to implicitly fuse individual features to generate global representation. For instance, Yu *et al.* [16] proposed an LSTM-based model to fuse the facial features, and Guo *et al.* [18] proposed a GNN model to exploit the emotional cues from the face, object, and skeleton features. These methods using implicit fusion assumed that there exists specific emotional interaction among individuals in a group [16], [17], [18]. The LSTM-based model requires the emotional interaction of individuals with the sequence relationship [16]. The GNN-based model considers that the emotional interaction graph is fully connected, which does not describe the actual individual emotional relationship in a group [18]. Although the methods using explicit fusion in GER do not need to assume relationship among the individuals [14], [18], [20], [33], [34], [35], it is still challenging to learn efficient fusion weights for the aggregation of individual features because fusion weight learning depends on global (scene) feature extraction, individual (facial) feature learning, and attention module utilization. In this work, we propose

TABLE I
DESCRIPTIONS OF THREE KINDS OF FEATURES IN OUR WORK

Type of Feature	Description
Individual (Facial) Emotion Feature	Generated by the FacesNet. The feature is extracted from a facial expression image in a group, which reflects the emotion state of an individual.
Global (Scene) Emotion Feature	Generated by the SceneNet. The feature is extracted from the whole group image, which contains the information of people, objects, and background.
BoW (Objects) Feature	The feature is the output of Bag-of-Word model which aggregates the semantic information of detected objects from a group.

to obtain efficient fusion weights by extracting multiscale global (scene) features to represent global emotion of a group, learning discriminative facial features to describe individual emotion, and designing an effective attention module to evaluate the correspondence of global and individual emotion.

C. Attention Mechanisms

In the neural network, the attention mechanism aims to mimic human brain actions of selectively concentrating on a few pieces of relevant information, while ignoring others. Up to now, various successful attention mechanisms were presented for different visual tasks. The review on attention mechanisms in [36] has divided them into four basic categories: channel attention [37] (what to pay attention to), spatial attention [38] (where to pay attention), temporal attention [39] (when to pay attention), and branch attention [40] (which to pay attention to). Each of these categories aims to learn the weights with the sigmoid function in the channel domain, the spatial domain, the temporal domain, and the features of the multibranch network, respectively.

In GER studies, most research adopted branch attention in feature aggregation. These GER methods used a multibranch network to extract and aggregate individual emotion features (facial features) in a group and learned the attention weights for individual features. In this work, we also use branch attention for individual features fusion but propose a new attention mechanism to learn the fusion weights more precisely for individual features.

D. Contrastive Learning

Contrastive learning is the technique to learn efficient feature representation by comparing different input samples. It aims at obtaining a small distance between the embeddings of two "similar" inputs (or a positive pair) and a large distance between the embeddings of two "dissimilar" inputs (or a negative pair). Recently, contrastive learning has been successfully used in computer vision, natural language processing (NLP), and other domains. Contrastive learning techniques can be divided into unsupervised and supervised. The unsupervised contrastive learning approaches, e.g., SimCLR [41], MoCo [42], and SimSiam [43], learn feature representations where the data itself provide supervision for downstream tasks [44]. The supervised contrastive learning approaches compare the positive and negative pairs indicated by their corresponding labels to learn more discriminative features for specific tasks [45], [46], [47].

For instance, the face recognition research in [58] used triplet contrastive learning to minimize the distance between an anchor and a positive sample and maximize the distance between the anchor and a negative sample. Literature review shows that contrastive learning is very effective for learning distinctive features and improving the performance of recognition networks. In this article, we use supervised contrastive learning in network training and design a new contrastive loss to constrain the training of the network to extract emotion-specific features for GER.

III. PROPOSED APPROACH

Our proposed method mainly comprises three subnetworks: FacesNet, SceneNet, and ObjectsNet. FacesNet is designed to extract and aggregate the individual emotion features (facial features) for GER from the faces in image. SceneNet is designed to capture the scene features (global features) and infer group emotion from the whole image. ObjectsNet is used to exploit the semantic information of detected objects to help determine group emotion. In addition, we use the prediction integration technique to fuse the outputs of FacesNet, SceneNet, and ObjectsNet to recognize group emotion. The overview of the proposed framework is shown in Fig. 2. The definitions of these three kinds of features in our work are listed in Table I.

A. FacesNet: Classification Using Facial Expression

The design of our FacesNet is shown in Fig. 3. Contrastive learning is used to help the network extract distinctive emotion features. A new attention mechanism named self-fusion module is designed to generate precise fusion weights for aggregation of individual facial features.

1) Face Detection and Backbone: All the faces are detected and cropped from the input image at the preprocessing stage. MTCNN [48], a high-performing face detector, is selected for face and landmark detection. Following the routines proposed in [18] and [20], we use the vgg16 network [49] pretrained with the vggface dataset [50] to predict the emotion of each person in the group. For the jth cropped face image in the ith image (I_{ij}^{f}), the output of the last fully connected layer in the vggface backbone represents the individual facial feature x_{ij}^{f}, and $\widehat{y}_{ij}$ denotes its predicted individual emotion.

2) Self-Fusion Module: Unlike other approaches that consider the emotion of each individual in the group to have the same contribution to GER, we propose a self-fusion module to learn ideal fusion weights to aggregate individual facial

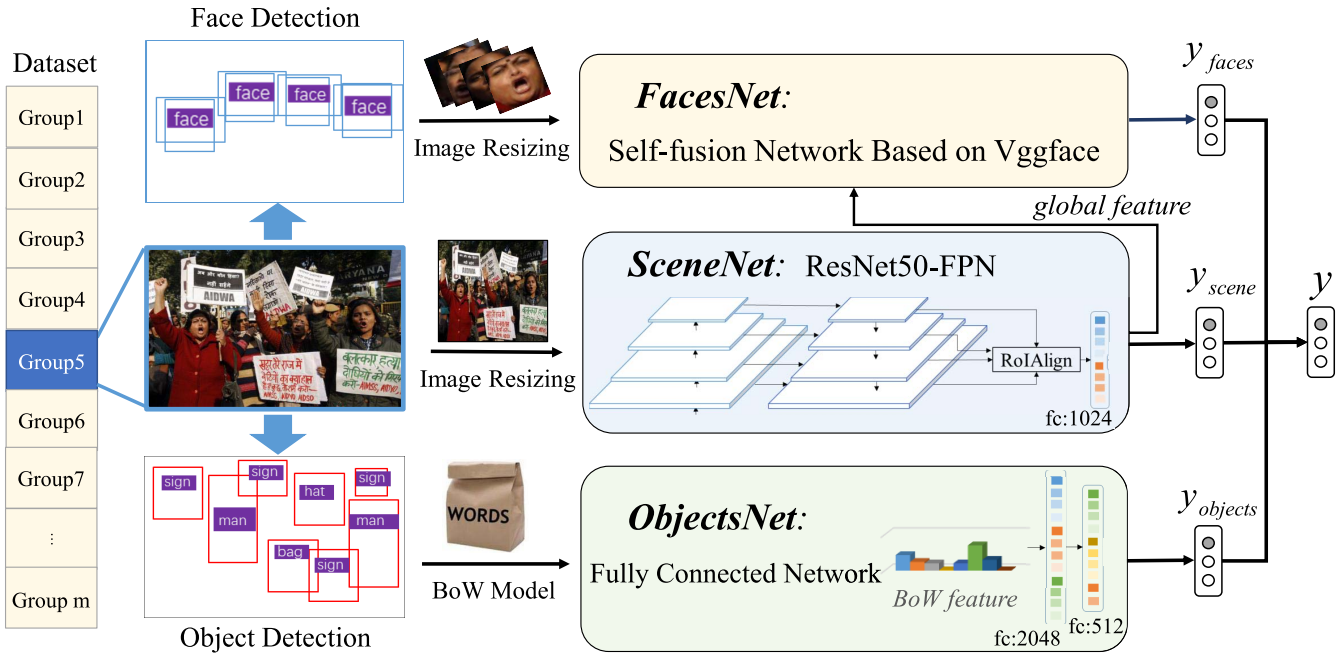


Fig. 2. Overview of the proposed framework for GER.

features to yield unique features for GER. Specifically, there are two objectives in the design of our self-fusion module. One is to characterize the global information and individual emotion with features. We use the trained SceneNet (will be mentioned in Section III-B) to extract the global information from the whole image and use the output of the last layer of SceneNet as the global emotion features. Similarly, we use the vggface network to learn the individual emotion feature from the facial expression and take the output of the last layer of vggface as the individual emotion feature. The other objective is to evaluate the similarity between the global and individual emotion features. Rather than constructing a discriminator to generate the weights for fusing individual facial features, we obtain the fusion weights by exploiting the cosine similarity of individual and global emotion features. Considering the dimension inconsistency in similarity calculation, we use fully connected layers with the ReLU activation function to separately map the global emotion features (scene features) and individual emotion features (facial features) into a common nonnegative subspace to ensure the obtained fusion weights fall in the range of [0, 1]. The fusion weights of the proposed self-fusion module is calculated as

$$v_{ij} = \frac{\omega_s\left(x_i^s\right) \cdot \omega_f\left(x_{ij}^f\right)}{\left\|\omega_s\left(x_i^s\right)\right\|_2 \left\|\omega_f\left(x_{ij}^f\right)\right\|_2} \tag{1}$$

where v_{ij} indicates the fusion weight of the jth individual facial feature in the ith group, and $\omega_s(\cdot)$ and $\omega_f(\cdot)$ denote two fully connected layers with ReLU activation function for the global emotion features and individual emotion features, respectively. $\omega_s(x_i^s) \cdot \omega_f(x_{ij}^f)$ denotes the inner product between the output vectors of $\omega_s(\cdot)$ and $\omega_f(\cdot)$, and $\|\cdot\|_2$ is the ℓ_2-norm.

The existing attention modules in the GER networks generate the attention weights for individuals by feeding global emotion features and individual facial features into a fully connected layer composed of one node with a sigmoid activation function [19], [20], [21]. The proposed self-fusion module is very different from the existing attention modules. As shown in Fig. 1, the sigmoid function in the existing attention modules likely traps important input values into a slow changing state and fails to describe the distinct importance of individual emotion. The proposed self-fusion module weights the individual features with the cosine similarity between the individual and global emotion features. Because it is sensitive to the change in input data, the proposed self-fusion module helps the network focus more on the individual features which are close to the global emotion features in terms of cosine similarity.

3) Loss Functions for FacesNet: Three loss functions are used for the backpropagation and updating of FacesNet (shown in Fig. 3) in the training process.

a) Cross-entropy loss for IER: The GER datasets do not provide labels for the emotion of individuals in the group. Following the routines of previous works [14], [19], we assign the individuals in the group emotion label and use the cross-entropy loss to constrain the backbone network. For the ith group, the cross-entropy loss for IER is formulated in the following equation:

$$L_{\text{IER}} = -\frac{1}{N_i}\sum_{j=1}^{N_i} \log \frac{e^{W_{y_{ij}}^{\mathrm{T}} x_{ij}^f}}{\sum_{k=1}^{C} e^{W_k^{\mathrm{T}} x_{ij}^f}} \tag{2}$$

where y_{ij} and x_{ij}^f denote the emotion label and facial feature of the jth individual in the ith group, respectively. W_k denotes the kth classifier for individual emotion. C is the number of

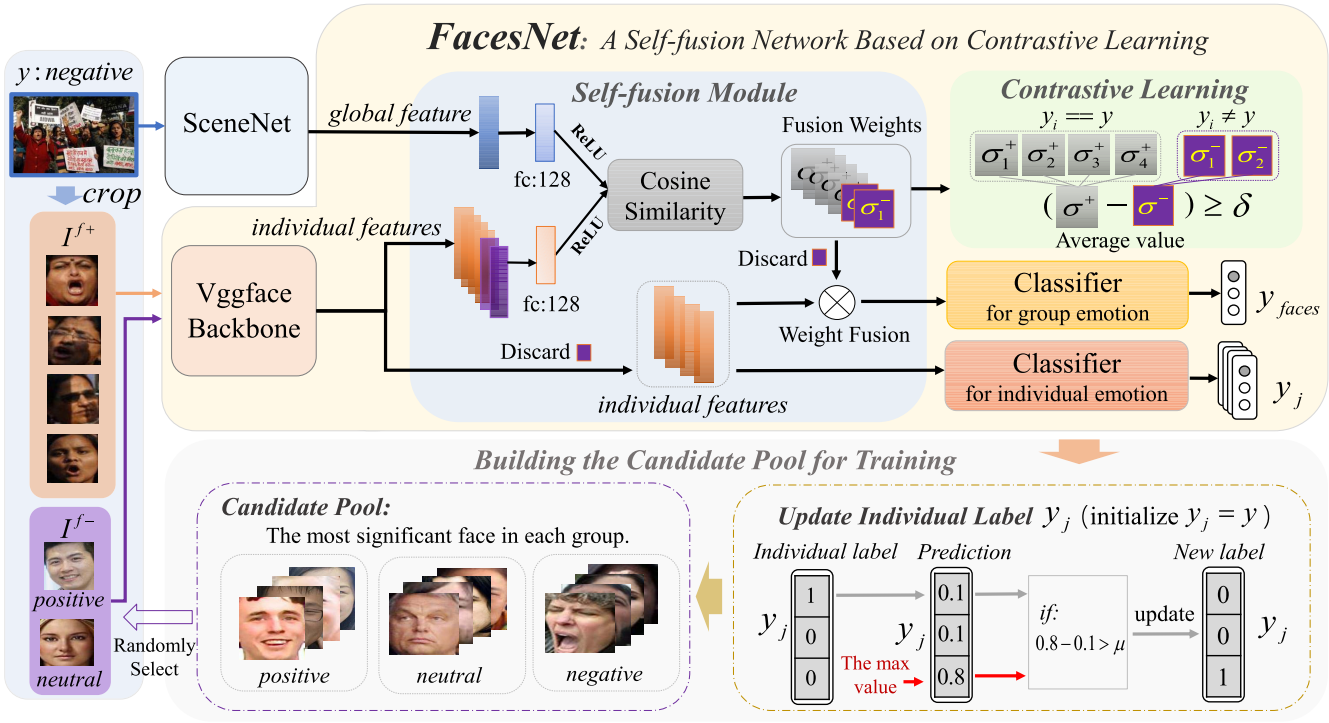


Fig. 3. Overview of the proposed FacesNet.

categories of individual emotion, and N_i is the number of faces in the ith image.

b) Contrastive learning loss: Fusion weights for individual facial features are generated by the self-fusion module. The values of fusion weights fall in the range of [0, 1]. To improve the distinguishability of individual facial features, we design a contrastive learning loss to enhance the difference between the learned weights of relevant and irrelevant individual emotions. Specifically, we divide individual face images into positive and negative sets. The former includes the individuals in a group who have the same emotion. The negative set includes individuals in a group whose emotions are not the same. Samples in the negative set are randomly selected from other groups with different emotion. To maximize the feature distance between positive and negative samples, we take the global (scene) features of the positive samples in the group as an anchor to calculate the fusion weights using (1) for positive samples and negative samples and propose a contrastive loss, as shown in the following equation:

$$L_{\mathrm{CL}} = \max\left\{0, \delta - \left(\sigma^{+} - \sigma^{-}\right)\right\}. \tag{3}$$

The symbol δ is a hyperparameter which indicates the least margin between the averaged weight of positive samples and negative samples. σ_i^{+} and σ_i^{-} denote the averaged weights of positive samples and negative samples in the ith positive–negative pair, respectively. They are calculated using the following equations:

$$\sigma_i^{+} = \frac{1}{N_i^{+}} \sum_{j \in \left\{\mathbf{x}_i^{f+}\right\}} \frac{\omega_s\left(x_i^s\right) \cdot \omega_f\left(x_{ij}^f\right)}{\left\|\omega_s\left(x_i^s\right)\right\|_2 \left\|\omega_f\left(x_{ij}^f\right)\right\|_2} \tag{4}$$

$$\sigma_i^{-} = \frac{1}{N_i^{-}} \sum_{j \in \left\{\mathbf{x}_i^{f-}\right\}} \frac{\omega_s\left(x_i^s\right) \cdot \omega_f\left(x_{ij}^f\right)}{\left\|\omega_s\left(x_i^s\right)\right\|_2 \left\|\omega_f\left(x_{ij}^f\right)\right\|_2}. \tag{5}$$

N_i^{+} and N_i^{-} denote the number of samples in the positive set $\{\mathbf{x}_i^{f+}\}$ and the negative set $\{\mathbf{x}_i^{f-}\}$ in the ith positive–negative pair, respectively. To ensure the effectiveness of the negative samples, we choose the face with maximum prediction probability in each group to construct a candidate pool and continually update the pool through the training process.

c) Cross-entropy loss for GER : To recognize the group emotion, we use the cross-entropy loss to constrain the group emotion output, which aggregates the recognized individual emotion by fusion weights. For the ith group, the cross-entropy loss for GER is formulated as

$$L_{\mathrm{GER}} = -\log \frac{e^{P_{y_i}^{\mathrm{T}} x_i^{\mathrm{agg}}}}{\sum_{k=1}^{C} e^{P_k^{\mathrm{T}} x_i^{\mathrm{agg}}}} \tag{6}$$

where P_k denotes the kth classifier for group emotion. x_i^{agg} is the fusion feature of the ith group. It is computed by weighted average of the individual facial features with the learned fusion weights using the following equation:

$$x_i^{\mathrm{agg}} = \frac{\sum_{j=1}^{N_i} v_{ij} x_{ij}^f}{\sum_{j=1}^{N_i} v_{ij}}. \tag{7}$$

d) Overall loss function: In the training stage of the proposed FacesNet, the above three losses are used simultaneously, and the overall loss function for the ith group sample is calculated as

$$\mathrm{Loss} = L_{\mathrm{IER}} + L_{\mathrm{CL}} + L_{\mathrm{GER}}. \tag{8}$$

4) Strategy for Updating Individual Label: Affective diversity has been used to describe the situation that members in the group may express different emotions in the same social event [51]. Inevitably, individual emotions from a group may be labeled incorrectly if the group emotion label is simply used to label all the individual emotions. An individual emotion relabeling strategy that minimizes the negative influence of unreliable labels was presented in [52]. Inspired by this work, we design a strategy to relabel the individual emotions in a group whose maximum prediction probabilities are significantly higher than their assigned group emotion label. The strategy for updating the individual emotion labels can be formulated as

$$y_{ij}^{*} = \begin{cases} l_{\max}, & \text{if } P_{\max} - P_{y_{ij}} > \mu \\ l_{y_{ij}}, & \text{otherwise} \end{cases} \tag{9}$$

where y_{ij}^{*} denotes the new label for the jth individual in the ith group, μ is a threshold, $P_{\max}$ is the maximum predicted probability, and $P_{y_{ij}}$ is the predicted probability of the given label y_{ij}. $l_{y_{ij}}$ and $l_{\max}$ are the original individual labels which are defined by group emotion and the index of the maximum prediction, respectively. Our FacesNet uses this label updating strategy after ten epochs and sets the threshold to 0.8.

B. SceneNet: Classification Using Scene Information

Scene feature contains the global information of the whole image, including humans, surrounding objects, and background, and it implies rich cues of group-level emotion. We design a subnetwork named SceneNet (shown in Fig. 2) to extract the global information from the whole image. SceneNet is composed of a ResNet50-FPN [53] and a RoIAlign function [54] that can extract global (scene) features in multiple scales. Two fully connected layers are used in SceneNet as the classifier for the GER task. The operation of scene feature extraction is defined as

$$x_i^{s} = \text{RoIAlign}(\varphi(I_i)) \tag{10}$$

where s is the index of the scene, I_i and x_i^{s} are the ith image and its corresponding global (scene) feature, respectively, and $\varphi(\cdot)$ denotes the network of ResNet50-FPN. RoIAlign$(\cdot)$ denotes the multiscale feature fusion function, which was proposed in Mask R-CNN [54] to properly align the extracted features with the input. Furthermore, the classifier based on global (scene) feature is defined as

$$\widehat{y}_{\text{scene}} = \omega\left(x^{s}\right) \tag{11}$$

where $\omega(\cdot)$ denotes the classifier function consisting of two fully connected layers, and $\widehat{y}_{\text{scene}}$ is the predicted group emotion based on the global (scene) feature. In the training process, the cross-entropy loss is used for the training of SceneNet.

C. ObjectsNet: Classification Using Object Features

ObjectsNet is developed to use the object information to infer the group emotion. It involves the feature extraction of the objects, feature aggregation, and group emotion classification. Different from previous research which extracted the object features using CNN, we use the detected semantic information as the object features to avoid intensive computation for extracting features for all the objects in the image.

To detect the objects in the group, we use a pretrained bottom-up [55] network to detect the objects in the image and assign a semantic label for each detected object. The object detector in [55] adopted a bottom-up mechanism (based on Faster R-CNN) to detect object regions and train on the MSCOCO dataset.

We use the BoW model to aggregate object semantic information into a BoW feature. We first count the frequency vector of the detected semantic labels for each image, and then use the algorithm of term frequency-inverse document frequency (TF-IDF) [56] to adjust the weights of object categories in the frequency vector. Finally, the adjusted frequency vector works as the BoW feature for group emotion classification.

To classify the group emotion based on the BoW feature of objects, we construct a classifier network with three fully connected layers (shown in Fig. 2). The dimension of the input layer is the same as the BoW feature (set to 1600). The number of the hidden nodes is set to 512, and the dimension of the output layer is the same as the number of emotion categories. Finally, the cross-entropy loss is used for backpropagation.

D. Predictions Integration

In this work, three separate predictions of group emotion are produced by FacesNet, SenceNet, and ObjectsNet. Predictions from these three networks are integrated into the final recognition result, which can be calculated as

$$\begin{aligned} \hat{y} &= \alpha\,\hat{y}_{\text{faces}} + \beta\,\hat{y}_{\text{scene}} + \gamma\,\hat{y}_{\text{objects}} \\ \text{s.t.} \quad & \alpha \geq 0, \beta \geq 0, \gamma \geq 0 \\ & \alpha + \beta + \gamma = 1 \end{aligned} \tag{12}$$

where α, β, and γ are the fusion parameters for the three predictions. We use the grid search approach to get the appropriate parameters for prediction integration.

IV. Experiments and Discussion

We conducted extensive experiments on three popular datasets, the Group Affective 2.0 (GAF2) [31] Dataset, the Group Affective 3.0 (GAF3) [32] Dataset, and the GroupEmoW [18] Dataset to evaluate the performance of the proposed framework. The numerical statistics of these three GER datasets are summarized in Table II. The elements listed in Table II are the number of images, and "-" represents "Not Applicable." Because the test sets of the GAF2 and GAF3 datasets were provided only for those who participated in the EmotiW competitions [31], [32], we used the training data to train our network and evaluated the classification performance using the validation data included in the GAF2 and GAF3 datasets. For the GroupEmoW dataset, we used the training data to train the model and used the test data to test the performance of our model. For all the three datasets, the maximum number of cropped faces from each image is set to 16, which is high enough for most images. We compared the proposed method with eight state-of-the-art baseline methods

TABLE II
STATISTICS OF THE THREE GER DATASETS. ELEMENTS LISTED IN THE TABLE ARE THE NUMBERS OF IMAGES, AND "-" REPRESENTS "NOT APPLICABLE"

Dataset / Partition	Group Affective 2.0				Group Affective 3.0				GroupEmoW			
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total
Train	1272	1199	1159	3630	3977	3080	2758	9815	4645	3463	3019	11127
Val	773	728	564	2065	1747	1368	1231	4346	1327	990	861	3178
Test	-	-	-	-	829	916	1266	3011	664	494	431	1589

TABLE III
COMPARISON OF CLASSIFICATION ACCURACIES FOR ALL CATEGORIES ON THE GROUPEMOW DATASET (IN %)

Sources / Methods		GAF 2.0 Validation			
		Positive	Neutral	Negative	Overall
Face+Scene	Surace+ [12]	68.61	59.63	76.05	67.75
	Abbas+ [13]	79.76	66.20	69.97	71.98
	Fujii+ [15]	75.68	69.64	77.33	74.22
	Fujii+ [20]	78.01	72.92	76.48	75.81
	Ours	**80.08**	**79.13**	**77.31**	**79.00**
Face+Scene+Object	Fujii+ [20]	**87.84**	77.55	74.10	**80.19**
	Ours	80.98	**80.39**	**76.01**	79.45

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACIES FOR ALL CATEGORIES ON THE GAF3 DATASET (IN %)

Sources / Methods		GAF 3.0 Validation			
		Positive	Neutral	Negative	Overall
Face+Scene	Fujii+ [15]	72.12	69.51	71.52	71.05
	Quach+ [57]	-	-	-	74.18
	Fujii+ [20]	78.42	71.19	**73.40**	74.34
	Ours	**86.61**	**78.36**	69.21	**79.08**
Face+Scene+Object	Fujii+ [20]	82.88	72.64	**74.32**	76.61
	Guo+ [18]	-	-	-	78.87
	Guo+ [58]	-	-	-	78.98
	Ours	**86.89**	**78.58**	70.35	**79.59**

TABLE V
COMPARISON OF CLASSIFICATION ACCURACIES FOR ALL CATEGORIES ON THE GROUPEMOW DATASET (IN %)

Sources / Methods		GroupEmoW Test			
		Positive	Neutral	Negative	Overall
Face+Scene	Khan+ [21]	-	-	-	89.36
	Ours	**94.58**	**85.63**	**86.77**	**89.68**
Face+Scene+Object	Guo+ [18]	-	-	-	89.93
	Khan+ [21]	-	-	-	89.61
	Ours	**94.58**	**85.46**	**88.63**	**90.06**

that performed the same three-class classification [12], [13], [15], [18], [20], [21], [57], [58].

A. Implementation Details

As shown in Fig. 2, the proposed method is composed of FacesNet, SceneNet, and ObjectsNet. Although the self-fusion module in FacesNet requires the global feature produced by the well-trained SceneNet, the above three networks are trained independently. To train FacesNet, we resized the cropped face images to 224 $\times$ 224 pixels, performed standard data augmentation (random horizontal flip, $\pm$20 rotation) to the resized face images, set the hyperparameter δ to 0.8 in contrastive learning loss, and used a stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 and back-propagation per four groups. For training SceneNet, we only trained the backbone of ResNet50-FPN, which was pretrained for image classification with ImageNet [59]. Specifically, we resized the input images to 800 $\times$ 800 pixels and used the optimizer of SGD with a learning rate of 0.001 and a batch size of 1. We froze the parameters of the backbone except RoIAlign and FC layers in the first epoch. For training ObjectsNet, we used the BoW feature as input which was constructed with objects' semantic information and used the SGD optimizer with a learning rate of 0.001 and a batch size of 512. We performed the optimization process for 50 epochs for all the networks. All the experiments were conducted on a Linux server with Intel Xeon CPU E5-2673 v4 2.30 GHz and GeForce GTX 2080Ti.

B. Comparison of Classification Performance

Tables III–V show the classification accuracies from the GAF2, GAF3, and GroupEmoW datasets, respectively. We compared the baseline methods under the same source scenario and show the accuracies on single and overall categories in the columns which are titled as "Positive," "Neutral," "Negative," and "Overall." "Face + Scene" denotes the classification results based on the cropped face and the whole image sources, and "Face + Scene + Object" denotes the results by simultaneously considering the cropped face, the whole image, and detected object sources.

As shown in Tables III–V, the proposed method outperformed the compared baseline methods with the usage of face and scene sources. The proposed method significantly improved the accuracy to 79.00% and 79.08% for the GAF2 and GAF3 datasets, respectively. The accuracy improvement is due to three main reasons: 1) feature pyramids are used to capture multiscale features, which generate more effective global features; 2) the proposed self-fusion module uses the metric of similarity to evaluate the correlation between global and individual emotion features and yields discriminative weights to aggregate individual features; and 3) the contrastive learning loss used in our model enhances the discriminative of learned features by differentiating positive

TABLE VI
COMPARISON OF THE OVERALL ACCURACY AND ACCURACIES FROM INDIVIDUAL SOURCES (IN %)

Datasets / Sources		Overall			Face			Scene			Object		
		pos	neu	neg	pos	neu	neg	pos	neu	neg	pos	neu	neg
Group Affect 2.0		acc=**79.45**			acc=76.83			acc=76.78			acc=61.36		
	pos	80.98	17.21	1.81	81.37	15.14	3.49	75.94	21.73	2.33	67.66	23.93	8.41
	neu	6.63	80.39	12.98	8.60	70.38	21.02	5.22	83.64	11.14	27.93	53.03	19.04
	neg	5.54	18.45	76.01	5.90	15.31	78.78	5.54	25.46	69.00	16.79	19.93	63.28
Group Affect 3.0		acc=**79.59**			acc=77.06			acc=77.20			acc=60.49		
	pos	86.89	10.19	2.92	87.06	7.90	5.04	82.31	14.83	2.86	71.15	17.80	11.05
	neu	8.48	78.58	12.94	10.60	69.15	20.25	7.31	81.14	11.55	33.70	46.86	19.44
	neg	8.29	21.36	70.35	8.77	19.58	71.65	8.29	26.16	65.56	22.91	16.57	60.52
GroupEmoW		acc=**90.06**			acc=89.30			acc=88.92			acc=75.52		
	pos	94.73	3.31	1.96	94.88	3.01	2.11	93.07	4.07	2.86	75.75	14.46	9.79
	neu	6.88	85.63	7.49	7.49	84.82	7.69	6.28	84.62	9.11	17.21	65.99	16.80
	neg	2.55	9.51	87.94	0.93	13.23	85.85	1.39	11.14	87.47	5.80	8.12	86.08

and negative samples. Comparing the results obtained with features from the faces and scene (indicated by "Face + Scene" in Tables III–V), the usage of object information benefited the performance of the proposed method, as indicated by "Face + Scene + Object." The overall results were also better than other baseline methods on the GAF3 and GroupEmoW datasets, which demonstrated that the proposed method can effectively exploit the relationship between object semantic information and group emotion. Although the baseline method in [20] achieved the best performance using all the resources on the GAF2 dataset, there is a difference in exploring the object information between the baseline method in [20] and our method. The method presented in [20] uses a CNN network to successively extract feature in detected object images. That requires high computational power for the GER system. Our method uses the BoW model to aggregate the detected object semantic information into a BoW feature. It dramatically decreases the computation requirements.

C. Ablation Study

We performed ablation studies to investigate how different sources, i.e., face, scene, and object, contribute to the overall classification performance, and how different modules contribute to FacesNet's performance. Tables VI and VII show the comparisons of using different sources and different modules, respectively.

1) Ablation Study on Different Sources: We conducted the recognition experiment with trained FacesNet, SceneNet, and ObjectsNet in different scenarios, including only one of the sources is available and all the sources are available. We calculated the recognition accuracy and confusion matrix for all the resources and each individual source to evaluate how much they contribute to the overall GER performance. The experiment results reported in Table VI show that facial features lead to more accurate recognition results than object features. This indicates that the facial information plays the most important role in recognizing group emotion. Similarly, the confusion matrix also shows Face outperformed other sources, i.e., the result obtained using Face consistently outperformed other sources for recognizing "positive" emotion. For results obtained using Object, although recognition accuracy was lower than those from other sources, the correct recognition of the "negative" class is the highest or second highest among all in the confusion matrix, which demonstrated the use of object information facilitated GER task especially for "negative" emotion. Table VI also shows that the confusion elements of "neu-neg" and "neg-neu" obtained by the source of Face or Scene are significantly larger than other confusion elements. It indicates that the "negative" and "neutral" emotions are easily confused in recognition.

2) Ablation Study on FacesNet: We performed an ablation study on three-module configurations with the vggface backbone to study the contribution of each of the three modules (self-fusion, contrastive loss, and label updating) in FacesNet. The first configuration includes only the proposed self-fusion module (vggface + self-fusion). The second configuration includes the self-fusion module and contrastive loss module (vggface + self-fusion + contrastive loss). The last one includes all the three modules (vggface + self-fusion + contrastive loss + update label).

For each configuration included in our experiments, we repeated the training process ten times and recorded the maximum accuracy on the validation sets of GAF 2, GAF 3, and GroupEmoW. We repeated the same experiments for the three configurations but replaced our self-fusion module with the attention module proposed in [19], [20], and [21] to evaluate the effectiveness of the self-fusion module. The mean accuracies of these evaluation experiments for all the three configurations are shown in Table VII. The mean accuracy increased as additional modules were added. This increase in accuracy proves that the performance of the proposed FacesNet benefits from contrastive learning and individual label updating strategy. The pairwise comparison shows that the proposed self-fusion module outperformed the attention module proposed in [19], [20], and [21] using the same experimental settings.

TABLE VII

COMPARISON OF THE MEAN CLASSIFICATION ACCURACIES (MEAN_V IN %) OBTAINED BY FACESNET WITH DIFFERENT MODULE COMBINATIONS. EXPERIMENTS WERE PERFORMED ON THE VALIDATION SETS OF GAF 2, GAF 3, AND GROUPEMOW

Module Combination	GAF2		GAF3		GroupEmoW	
	Avg_V	*t*-test	Avg_V	*t*-test	Avg_V	*t*-test
vggface + attetnion	72.04	*t(9)*=-5.26, *p*=.0005211	73.14	*t(9)*=-5.72, *p*=.0002859	86.86	*t(9)*=-15.36, *p*=.0000001
vggface + self-fusion	74.16		74.51		88.36	
vggface + attention + contrastive learning	72.72	*t(9)*=-8.17, *p*=.0000187	73.24	*t(9)*=-11.37, *p*=.0000012	87.09	*t(9)*=-13.28, *p*=.0000003
vggface + self-fusion + contrastive learning	74.82		75.83		88.69	
vggface + attention + contrastive learning + update label	73.92	*t(9)*=-13.33, *p*=.0000003	73.95	*t(9)*=-15.11, *p*=.0000001	87.42	*t(9)*=-14.92, *p*=.0000001
vggface + self-fusion + contrastive learning + update label (ours)	**76.66**		**76.79**		**88.90**	

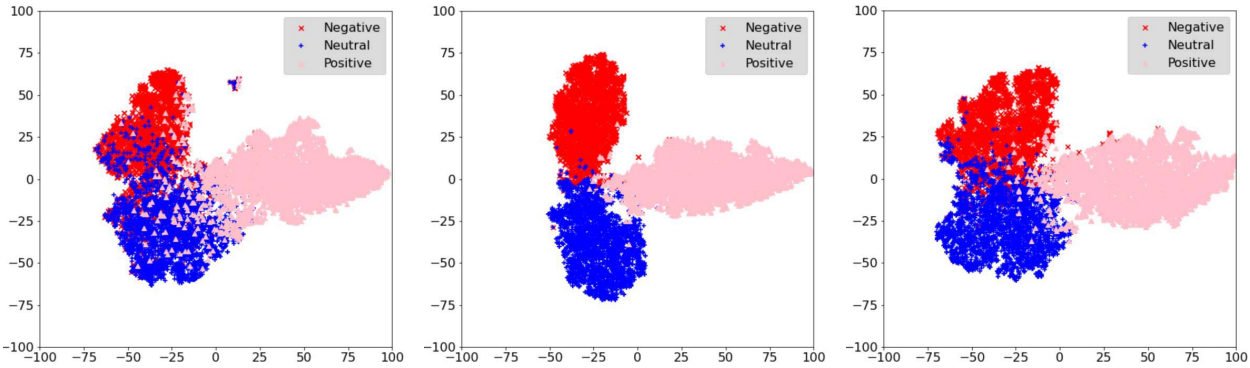


Fig. 4. *t*-SNE visualization of fusion feature distributions of the proposed FacesNet on the GAF3 dataset. (a) Cross-entropy loss with the self-fusion module. (b) Contrastive loss and cross-entropy loss with the self-fusion module. (c) Contrastive loss and cross-entropy loss with the attention module.

We also performed the paired *t*-test [60] on accuracies of our self-fusion module and the attention module proposed in [19], [20], and [21]. As the paired *t*-test relies heavily on the assumption of normal distribution, we first used the Shapiro–Wilk test to verify the distribution of classification accuracies for each model. The results of the Shapiro–Wilk test indicate the obtained classification accuracies follow normal distribution. Then, we calculated a test statistic t for the difference between the true mean accuracy for the network with the attention module and that for the network with the self-fusion module, using a 0.05 significance level. The results are also reported in Table VII. We made a hypothesis that the mean accuracy obtained using the attention module might be higher than using the self-fusion module. However, the significance level p was less than 0.05, which means our hypothesis is false. In other words, the proposed self-fusion module yielded better performance.

D. Visualization Results

To evaluate the contribution of contrastive learning and self-fusion in FacesNet, we use the *t*-SNE algorithm [61] to visualize the distribution of learned features. *t*-SNE is one of the popular feature dimension reduction methods used to visualize the features learned by deep learning models on a 2-D plane. We use *t*-SNE to visualize the representation features obtained by fusing individual features. We chose the learned feature of the training set with the highest recognition accuracy on the GAF3 and GroupEmoW datasets for *t*-SNE visualization experiments. Figs. 4 and 5 show the distributions of the learned features with and without contrastive learning for the GAF3 (see Fig. 4) and GroupEmoW (see Fig. 5) datasets.

Figs. 4(a) and 5(a) show the distributions of the learned features using only the cross-entropy loss for our self-fusion module. Figs. 4(b) and 5(b) show the distributions of the learned features using both the cross-entropy and contrastive losses for our self-fusion module. The learned features with contrastive learning clearly have larger inter-class distance and clearer boundaries of feature distribution compared with the learned features without contrastive learning.

We also included the distributions of the learned features using the previous attention module [19], [20], [21] to replace our self-fusion module in FacesNet for the two datasets in Figs. 4(c) and 5(c) for comparison. The difference between two sets of distributions [see Figs. 4(b) and (c) and 5(b) and (c)] show that our self-fusion module produced larger separation

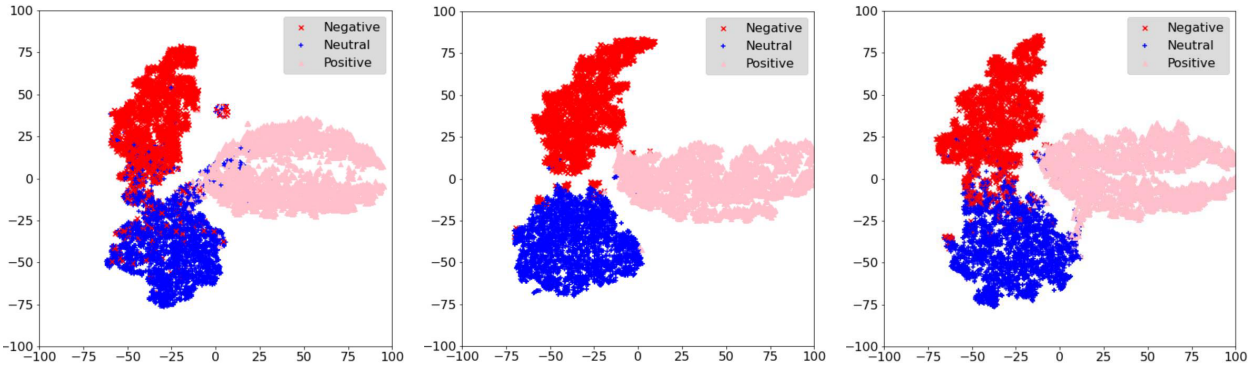


Fig. 5. t-SNE visualization of fusion feature distributions of the proposed FacesNet on the GroupEmoW dataset. (a) Cross-entropy loss with the self-fusion module. (b) Contrastive loss and cross-entropy loss with the self-fusion module. (c) Contrastive loss and cross-entropy loss with the attention module.

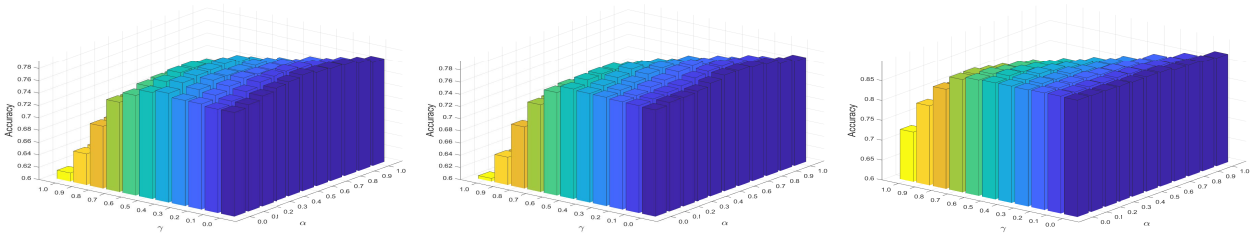


Fig. 6. Accuracies with different α and γ on the datasets of (a) GAF2, (b) GAF3, and (c) GroupEmoW.

and better feature representation for classification than the attention module.

E. Effects of Integration Parameters

We used averaged weights to fuse the recognition results from FacesNet, SceneNet, and ObjectsNet. To achieve the best performance of our work, we used a grid search technique to obtain the near-optimal hyperparameters α, β, and γ as mentioned in Section III-D. Due to the constraints of $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$, the hyperparameter β can be expressed as $1 - \alpha - \gamma$. We further tested the impact of the values of α and γ, by adjusting them one at a time. More specifically, we varied the parameter α in the range of [0, 1] and γ in the range of $[\alpha, 1]$ with a stride of 0.01. The integrated results tested on different $\{\alpha, \gamma\}$-pair values are shown in Fig. 6. The best performances on the GAF2, GAF3 and GroupEmoW datasets were achieved by setting $\alpha = 0.35, \beta = 0.2, \gamma = 0.45$, $\alpha = 0.45, \beta = 0.2, \gamma = 0.35$, and $\alpha = 0.37, \beta = 0.15, \gamma = 0.48$, respectively. These results indicate that of the three networks, FacesNet contributed the most to the recognition performance.

V. Conclusion

In this work, we propose an effective model for GER by exploring the emotion information from the facial expression of people in a group, the whole image, and the object semantic information. Specifically, we first propose a new contrastive learning-based self-fusion network called FacesNet to fuse the individual facial features together to generate more discriminative features for GER. Second, we combine the feature pyramid network with a ResNet50 backbone to capture the global (scene) features in multiple scales. We then use the BoW model to aggregate the semantic information of the detected objects and construct a fully connected network to classify the learned BoW features. We linearly integrate the outputs of these three networks as the final output of this unique framework for GER. We compared the proposed method with eight state-of-the-art baseline methods to demonstrate its effectiveness.

References

[1] I. T. Meftah, N. Le Thanh, and C. B. Amar, "Detecting depression using multimodal approach of emotion recognition," in *Proc. IEEE Int. Conf. Complex Syst. (ICCS)*, Nov. 2012, pp. 1–6.

[2] M. Nabi, H. Mousavi, H. Rabiee, M. Ravanbakhsh, V. Murino, and N. Sebe, "Abnormal event recognition in crowd environments," in *Applied Cloud Deep Semantic Recognition*. Boca Raton, FL, USA: Auerbach, 2018, pp. 37–56.

[3] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[4] T. Ni, C. Zhang, and X. Gu, "Transfer model collaborating metric learning and dictionary learning for cross-domain facial expression recognition," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1213–1222, Oct. 2021.

[5] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Aug. 2016.

[6] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018.

[7] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[8] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Exp. Syst. Appl.*, vol. 41, no. 13, pp. 5858–5869, Oct. 2014.

[9] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.

[10] L. Shu *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[11] M. Bradley and P. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, Mar. 2000.

[12] L. Surace, M. Patacchiola, E. B. Sönmez, W. Spataro, and A. Cangelosi, "Emotion recognition in the wild using deep neural networks and Bayesian classifiers," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 593–597.

[13] A. Abbas and S. K. Chalup, "Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 561–568.

[14] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group emotion recognition with individual facial emotion CNNs and global image based CNNs," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 549–552.

[15] K. Fujii, D. Sugimura, and T. Hamamoto, "Hierarchical group-level emotion recognition in the wild," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[16] D. Yu, L. Xingyu, D. Shuzhan, and Y. Lei, "Group emotion recognition based on global and local features," *IEEE Access*, vol. 7, pp. 111617–111624, 2019.

[17] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "LSTM for dynamic emotion and group emotion recognition in the wild," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 451–457.

[18] X. Guo, L. F. Polania, B. Zhu, C. Boncelet, and K. E. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2921–2930.

[19] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli, "An attention model for group-level emotion recognition," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 611–615.

[20] K. Fujii, D. Sugimura, and T. Hamamoto, "Hierarchical group-level emotion recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3892–3906, 2021.

[21] A. S. Khan, Z. Li, J. Cai, and Y. Tong, "Regional attention networks with context-aware fusion for group emotion recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1150–1159.

[22] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 1470.

[23] T. Zhang, X. Gong, and C. L. P. Chen, "BMT-Net: Broad multitask transformer network for sentiment analysis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6232–6243, Jul. 2022.

[24] T. Zhang, X. Wang, X. Xu, and C. P. Chen, "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Cybern.*, vol. 13, no. 1, pp. 379–388, Jan./Mar. 2022.

[25] X. Gong, T. Zhang, C. L. P. Chen, and Z. Liu, "Research review for broad learning system: Algorithms, theory, and applications," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 8922–8950, Sep. 2022.

[26] T. Zhang *et al.*, "Cross-database micro-expression recognition: A benchmark," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 544–559, Feb. 2022.

[27] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 503–510.

[28] C. Wang, S. Wang, and G. Liang, "Identity- and pose-robust facial expression recognition through adversarial feature learning," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 238–246.

[29] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812.

[30] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–8.

[31] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: EmotiW 5.0," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 524–528.

[32] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "EmotiW 2018: Audio-video, Student engagement and group-level affect prediction," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 653–656.

[33] B. Balaji and V. R. M. Oruganti, "Multi-level feature fusion for group-level emotion recognition," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 583–586.

[34] X. Guo, L. F. Polanía, and K. E. Barner, "Group-level emotion recognition using deep models on image scene, faces, and skeletons," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 603–608.

[35] K. Wang *et al.*, "Cascade attention networks for group emotion recognition with face, body and image cues," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 640–645.

[36] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," 2021, *arXiv:2111.07624*.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[39] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial–temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4733–4742.

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[43] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[44] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[45] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[46] M. Zheng *et al.*, "Weakly supervised contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10042–10051.

[47] X. Zhang, M. Xu, X. Zhou, and G. Guo, "Supervised contrastive learning for facial kinship recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–5.

[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[50] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.* Durham, U.K.: British Machine Vision Association, 2015, pp. 1–12.

[51] S. G. Barsade and A. P. Knight, "Group affect," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 2, no. 1, pp. 21–46, 2015.

[52] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6897–6906.

[53] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[54] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[55] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

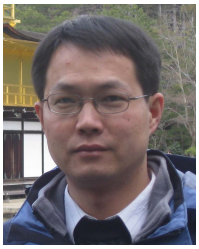
[56] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, 2003, vol. 242, no. 1, pp. 29–48.

[57] K. G. Quach, N. Le, C. N. Duong, I. Jalata, K. Roy, and K. Luu, "Non-volume preserving-based fusion to group-level emotion recognition on crowd videos," 2018, *arXiv:1811.11849*.

[58] X. Guo, B. Zhu, L. F. Polanía, C. Boncelet, and K. E. Barner, "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 635–639.
[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
[60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.
[61] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

Xingzhi Wang received the B.S. degree from the Guangdong University of Finance and Economics, Guangzhou, China, in 2016, and the M.S. degree from Huaqiao University, Quanzhou, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Electrics and Information Technology, Sun Yat-sen University, Guangzhou.

His research interests include emotion recognition and computer vision.

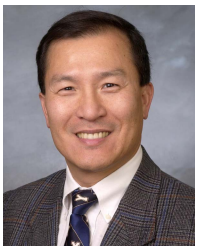


Dong Zhang received the B.S.E.E. and M.S. degrees from Nanjing University, Nanjing, China, in 1999 and 2003, respectively, and the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2009.

He is currently an Associate Professor with the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, pattern recognition, and information hiding.

Hong-Zhou Tan (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, and the South China University of Technology, Guangzhou, China, in 1998.

From 1998 to 2004, he was with several universities and IT companies in Hong Kong, Singapore, and Canada. Since 2004, he has been a Full Professor with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou. His current research interests include the Internet of Things and pattern recognition.



Dah-Jye Lee received the M.S. and Ph.D. degrees in electrical engineering from Texas Tech University, Lubbock, TX, USA, in 1987 and 1990, respectively, and the M.B.A. degree from Shenandoah University, Winchester, VA, USA, in 1999.

He worked in the machine vision industry for 11 years before joining Brigham Young University (BYU), Provo, UT, USA, in 2001. He is currently a Professor with the Department of Electrical and Computer Engineering, BYU. His research interests include artificial intelligence, robotic vision, high-performance visual computing, and visual inspection automation.