IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 12, 2020, accepted December 29, 2020, date of publication January 1, 2021,
date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048693

Graph Reasoning-Based Emotion Recognition Network

QINQUAN GAO[1,2], HANXIN ZENG[1], GEN LI[2], AND TONG TONG[1,2]
[1]College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China
[2]Imperial Vision Technology, Fuzhou 350002, China

Corresponding author: Tong Tong (ttraveltong@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61901120 and Grant 61802065, and in part by the Science and Technology Program of Fujian Province of China under Grant 2019YZ016006.

ABSTRACT Semantic information from images can be used to improve the performance of deep learning methods in recognizing human emotions. In this paper, we propose a novel framework based on the graph convolutional network for emotion recognition by utilizing the semantic relationships of different regions. First, we extract the salient image regions within video frame clips by using the bottom-up attention module to construct the node features of a graph. Then, we build the graphs containing the node features and the semantic correlations of nodes by using the graph convolutional network. For refinement, each node feature of graph vectors is enhanced via a gated recurrent unit consisting of gate and memory units to remove redundant feature information. Experimental results show that our proposed method achieves superior performance over state-of-the-art approaches for the emotion recognition on the CEAR and AFEW datasets.

INDEX TERMS Emotion recognition, graph convolutional neural networks, contextual spatiotemporal features.

I. INTRODUCTION

Human emotions substantially affect the mutual communication and decision making of humans in daily life [1]. Nowadays, recognizing human emotions plays an increasingly important role in various applications. The ability of intelligent service robots to recognize the emotions of an interactive user is indispensable. In social media platforms, extracting social sentiment from textual data is beneficial for digital monitoring and online information pushing. In the medical field, an emotion recognition model can recognize the emotions of patients and provide appropriate treatments by analyzing physiological signals. Unlike object recognition and classification, emotion recognition in real life requires sensory-based reasoning. For example, people deduce the emotion category by subconsciously reasoning the facial expressions, voice intonations, and body movements of human beings.

The efforts in emotion recognition have mostly been divided into two main categories: physiological signal-based methods [2] and non-physiological signal-based methods. Physiological signals include electroencephalogram (EEG) signal [3]–[5], galvanic skin response signal [6]–[8], and electrocardiogram signal [9], [10].Other methods not based on physiological signals identify emotions by extracting the features from various types of data, such as videos/images, speeches, and texts. Traditional methods based on the videos/images recognize emotions by utilizing the handcrafted features [11]–[13]. Handcrafted features are extracted to form vectors representing the geometry of the face, which include the facial shape and locations. Shan *et al.* [11] used support vector machine (SVM) classifiers to recognize facial emotions with local binary pattern (LBP) representation. Zhao and Pietikainen [12] recognized the dynamic texture of facial expressions by extracting facial local information and its spatial locations using volume local binary patterns (VLBP) operator. However, handcrafted feature-based methods normally require precise and reliable detection and tracking of facial components, which are difficult in many situations. With the proposal of deep learning methods, the above problems can be solved.

In recent years, Convolutional Neural Network (CNN)-based methods [14]–[16] have achieved more accurate and robust emotion recognition than previous methods with changes in surrounding information. Yu and Zhang [15] created a model to recognize the emotions of static images,

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

which contains three face detectors and a multiple deep CNNs module. Unlike traditional algorithms that can only detect facial emotions with frontal facial components, this recognition system can detect spontaneous facial expression in the wild. In visual emotion recognition, other visual cues such as body gestures, actions, and environmental contexts can show additional useful information. Thus, Lee *et al.* [14] integrated the facial expressions and surrounding information of people with adaptive fusion networks to demonstrate that the performance of emotion recognition networks can be remarkably boosted by integrating facial and context information. In fact, the emotion recognition system can analyze features at the local pixel level, which are extracted by a specific convolution receptive field. However, learning the relationships of high-level semantic information among regions is difficult.

In the topology data structures, the Graph Convolutional Network (GCN) can fully extract the relationship features between node vectors. Inspired by the application of GCN in EEG emotion recognition and textual emotion recognition, we present a novel GCN-based framework to recognize human emotions by globally capturing the relationships between different semantic regions, which included face and scene contexts. First, we identify salient regions in videos at the object level with bottom-up attention [17], which is achieved by using Faster R-CNN [18] and is similar to the human visual system. Then, we convert multiple regions of video frames into corresponding node vectors of the graph structure. In this way, the data conversion between Euclidean structure and graph structures can be achieved. We also generate semantic relationships of node vectors by utilizing GCN to explore the interrelation between these node vectors of notable regions. Although diverse semantic relationships could contribute to boosting the accuracy of emotion recognition, part of the relationships among them are redundant. Thus, we utilize a gated recurrent unit (GRU) to select high-level graph features before generating the final representation of the entire video frame. In specific, our proposed GCN-based emotion recognition framework has the following major contributions:

1) Unlike the traditional CNN-based methods that simply consider the context of images or video frame input, our proposed network recognizes emotions from context-aware emotion datasets with reasonable and filter mechanism, which can reason the relationships of different regions and remove redundant information among node vectors.
2) This work examines distinct approaches to verify that the proposed framework extracts spatiotemporal features more efficiently than previous methods over the Context-aware Emotion Recognition (CAER) [14] and Acted Facial Expressions in the Wild (AFEW) [19] datasets, which includes the complex context information.
3) The proposed method is more accurate than prior methods [14], [20], [21] for emotion recognition on the CAER and AFEW datasets. Our network outperforms these baseline methods by 7%-9% improvement on the CAER dataset. Furthermore, we conduct ablation studies to justify the effectiveness of combining GCN and GRU.

II. RELATED WORK

A. TRADITIONAL METHODS FOR EMOTION RECOGNITION

Before the popularity of CNN-based methods, most emotion recognition research has been dominated by traditional methods using handcrafted features or shallow classifiers such as the Facial Action Coding System (FACS) with action units (AUs) [22], [23], local binary patterns (LBPs) [11], [12], [24], and sparse learning [13]. Tian *et al.* [22] developed an Automatic Face Analysis (AFA) system to recognize emotions by describing slight changes in a face into AUs of FASC, which contains permanent and transitory facial features. Different from some studies based on original face images, some methods recognize facial expressions using statistical local facial features. Shan *et al.* [11] extracted Boosted-LBP features that represent the most salient LBP features and then utilized SVM classifiers to perform emotion recognition using the extracted the Boosted-LBP features. This method can also work well for low-resolution facial expression. However, Zhong *et al.* [13] observed in facial expression recognition that only a few facial components are useful. In [13], a framework of two-stage multi-task sparse learning was proposed to efficiently locate common patches of each expression and learn particular expression patches, and then emotional classification results were presented by trained SVM classifiers with the facial patches. Wang *et al.* [23] exploited the complicated semantic relationships among facial AUs to recognize the facial emotions generated by the restricted Boltzmann machine. Although the above-mentioned traditional methods have shown excellent emotion recognition results by extracting the facial features on the datasets generated in lab-controlled environments, the robustness of these methods is unsatisfactory when the face images include various head pose changes, which may make the traditional methods fail to extract the useful AUs or LBP features.

B. DEEP LEARNING METHODS FOR EMOTION RECOGNITION

With the rapid development of CNN in emotion recognition, the approaches of extracting facial features have been injected into the new vigor and vitality. CNN-based methods have achieved more robust and superior performance than traditional methods in the wild datasets by extracting high-level semantic features [25]. Liu *et al.* [26] presented a unified framework, Boosted Deep Belief Network, for integrating three training stages through joint training. The method based on joint training strengthens the capabilities of facial feature selection and facial expression classification. Based on the above methods, these facial expression systems detect emotions through various spatial facial features. Considering

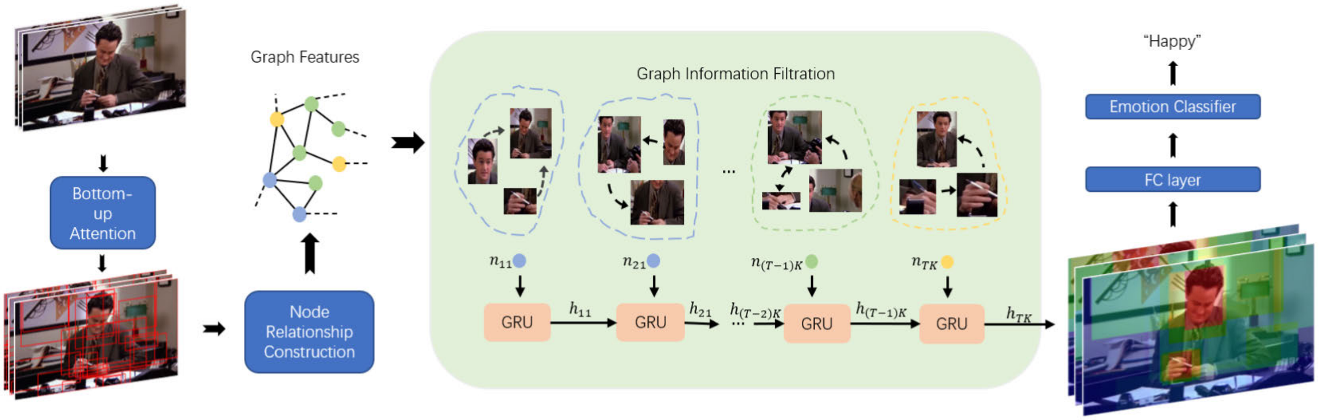


FIGURE 1. The overview of Graph Reasoning-based emotion recognition network (GRERN).

temporal information, Jung *et al.* [27] proposed a deep network that includes two models combined with a new integration method to extract temporal features from image sequences and facial landmark points. The audio stream information also contributes to the emotion recognition of videos in addition to classical facial expression features. Kahou *et al.* [28] developed a framework to merge four modality models. The first model based on several convolutional networks is trained to recognize emotion categories by extracting the facial features in each video frame. Then, another model captures emotion features from audio stream. The third model captures body actions by analyzing the temporal features of video frames. The final model is a shallow network trained to learn the mouth action features in the video sequence. Although this method highlights the importance of different data types for emotion recognition, the system only works on the frontal face. Thus, the robustness of this emotion recognition system is limited in the wild environment dataset. As described above, Zhang *et al.* [29] proposed a framework to recognize pose-variant facial expression and body pose. In addition, this network introduced the generative adversarial network to enlarge and enrich the training dataset by generating facial expression images under different poses. However, those methods analyze the emotion features from the facial expression dataset without exploiting the contextual information of the environment. Lee *et al.* [14] provided a dataset collected from 79 TV shows. This dataset contains not only the facial expression but also complex and real surrounding contextual information. Lee *et al.* used this dataset to evaluate the role of contextual information in emotion recognition. They proposed a CNN-based framework to detect emotions by fusing the facial features and contextual features. Compared with methods focusing only on facial expression features, such a method has achieved better accuracy of emotion recognition.

C. GRAPH CONVOLUTION NETWORK

Recently, GCN has been widely used for emotion recognition, especially in physiological signals and text emotion recognition, because of its powerful capacity for extracting the relationships of emotion features [30]. GCN can exploit the relationship features between graph structure data, which cannot be achieved by CNNs. Therefore, some studies based on GCN have been proposed to effectively tackle tasks that require rich relational structure data and depend on the global information of graph data to achieve their goals, such as skeleton action recognition [31], natural language processing [32], and emotion recognition [20], [33]–[36]. However, most GCN-based emotion recognition methods use texts or physiological signals. Zhang *et al.* [20] established a two-branch network for image emotion recognition. One branch utilizes contextual features to deduce emotion information through GCN. The other one learns body action features by training the VGG-16 [37] network. This method was implemented on the image dataset and utilized the context relationships on the spatial domain. However, the research on the combination of the spatial and temporal contextual features in video emotion recognition is still a daunting challenge. In this paper, we propose a network integrating the GCN and GRU to recognize human emotions in the videos, which can consider the contextual factors from video sequence to infer emotion features in the surrounding.

III. PROPOSED METHOD

This section describes the detailed structure of the Graph Reasoning-based Emotion Recognition Network (GRERN) for videos, as shown in Figure 1. Our approach extracts spatiotemporal features of environmental context and facial expression for emotion recognition. However, GCN is limited to directly process Euclidean structure data-like video frames. Therefore, in the first stage, we convert the regions of video frame to node vectors of graph structure (a set of video frames corresponds to a graph structure) by the bottom-up attention model [17] (Sec. III-A). In this way, we map the consecutive video frames into the topology structure space to match the next GCN operation. Then, we establish connections between these node vectors of the graph with learnable weights. In the second stage of learning semantic relationship information among the node features, we apply four layers of GCN after the first stage of operation, which can extract the spatial features and the temporal information of the surrounding context (Sec. III-B). Then, we convert a set of video frames to a graph structure. We use a GRU to refine the graph features by removing redundant features. Finally, we input the graph features into a SoftMax layer to classify the emotion categories (Sec. III-C).

A. GRAPH NODE STRUCTURE TRANSITION

In the beginning, we denote a set of T video frames of the CAER datasets as $V = \{v_1, \ldots, v_T\}$. To obtain the graph structure data from videos, we transform video frames V into the graph structure $G = \{N_1, \ldots, N_T\}, N_i \in \mathbb{R}^{K \times D}$ through the bottom-up attention mechanism. This graph structure is composed of multiple node vector sets. In other word, we convert the i video frame v_i into the i set of node vectors $N_i = \{n_{i1}, \ldots, n_{iK}\}, n_{ij} \in \mathbb{R}^D$, and each single set of node vectors contains K node vectors. In addition, the single node vector is D-dimensional. We utilize the bottom-up attention network to implement the conversion process with Faster R-CNN [18]. To effectively extract the sets of node vectors G for subsequent semantic reasoning, we detect the discriminative regions of video frames with non-maximum suppression, and set the 0.7 IoU threshold to select the salient regions. In addition, we set the confidence threshold of 0.3 to remove some regions with class probability lower than the threshold. This attention network selects the top K regions of the video frame v_i and outputs the corresponding feature vectors $X_i = \{x_{i1}, \ldots, x_{iK}\}$, which are ranked according to the detection confidence scores. In this way, each video frame is represented by K node vectors. Finally, we utilize a fully connected layer to transform the feature vector x_{ij} to a D-dimensional node vector n_{ij}, and then concatenate all node feature vectors to compose the final graph structure G as the following equation:

$$n_{ij} = W_f x_{ij} + b_f \tag{1}$$

where W_f and b_f are parameters of the fully connected layer. x_{ij} is the feature vector of the detection model, and n_{ij} is the corresponding node vector. In summary, we form a set of node vectors $N_i = \{n_{i1}, \ldots, n_{iK}\}$ with K node vectors, which represents a video frame. Futhermore, $G = \{N_1, \ldots, N_T\}$ represents each video clip that consists of T video frames.

B. GRAPH RELATIONSHIP REASONING

In the Fourier domain, a GCN model can extract the features between the node vectors. Therefore, we apply the GCN model to extract the emotional information hidden in the node vectors. To start with node vectors N as well as an adjacency matrix A, a multi-layer GCN model can be expressed as follows:

$$H^{l+1} = ReLu(\hat{D}^{-\frac{1}{2}}\tilde{A}\hat{D}^{-\frac{1}{2}}H^l W^l) \tag{2}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix A with added an identity matrix I_N. $\hat{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix. W^l is l^{th} layer learnable weights, and H^l is the output after l^{th} activation layer.

To illustrate the effect of the adjacent matrix, we define $\hat{D}^{-\frac{1}{2}}\tilde{A}\hat{D}^{-\frac{1}{2}}$ as $\hat{A}$. The adjacent matrix $\hat{A}$ represent the interaction intensity between each pair of node vectors. The high element value $\hat{A}_{ij}$ of the adjacent matrix means that the relationship between n_i and n_j node vectors is strongly correlated. In [30], the adjacent matrix $\hat{A}$ is calculated before the convolutional operation. To consider the changes in the relationship between the node vectors in the GCN propagation, we convert the adjacent matrix $\hat{A}$ to be a trainable matrix in accordance with the following rule:

$$\hat{A}(n_i, n_j) = \Theta(n_i)^T \Psi(n_j) \tag{3}$$

where n_i and n_j denotes the node vectors of the graph. $\Theta(v_i)$ and $\Psi(v_j)$ are two trainable vectors calculated with weights:

$$\Theta(v_i) = W_\theta v_i, \Psi(v_j) = W_\psi v_j \tag{4}$$

In the reasoning stage, we use GCN to transmit the emotional information of nodes through the adjacent matrix. Finally, we introduce residual connections into each GCN layer:

$$G^{l+1} = ReLu(W_r(\hat{A}G^l W^l) + G^l) \tag{5}$$

where W^l denotes the parameters of the l^{th} GCN layer and W_r is the weight of residual connection. We apply a ReLU activation function after each layer of GCN. In this way, we realize four times such well-defined GCN operations on the node vectors of the graph to extract spatial-temporal emotion features effectively. In particular, to recognize emotions in a single image, we can set the number T of video frames to 1.

C. GRAPH INFORMATION FILTRATION

In this section, we refine graph features and extract the discriminative features of the graph by removing the redundant features to obtain the final emotional features. In specific, we input the node vectors of graph G one by one into the GRU layer to capture the long-term information and remove the redundant information. As shown in Figure 2, the GRU adaptively captures the emotion features through different gates, such as update gate u_{ij} and reset gate r_{ij}. The update gate u_{ij} controls how much information is transferred from the previous state to the hidden state. The reset gate r_{ij} effectively controls the hidden state to ignore the information from the previous state which is irrelevant to the current state.

The presentation of the updated memory h_{ij} is a linear interpolation between the hidden state $\hat{h_{ij}}$ and the previous state $h_{(i-1)j}/h_{T(j-1)}$ based on the update gate u_{ij} as:

$$h_{ij} = \begin{cases} u_{ij} \circ \hat{h}_{ij} + (1 - u_{ij}) \circ h_{(i-1)j}, & i \neq 1 \\ u_{ij} \circ \hat{h}_{ij} + (1 - u_{ij}) \circ h_{T(j-1)}, & i = 1 \end{cases} \tag{6}$$

where $\circ$ is an element-wise multiplication. The update gate $u_{ij} \in [0, 1]$ decides the degree of the unit updated, which is

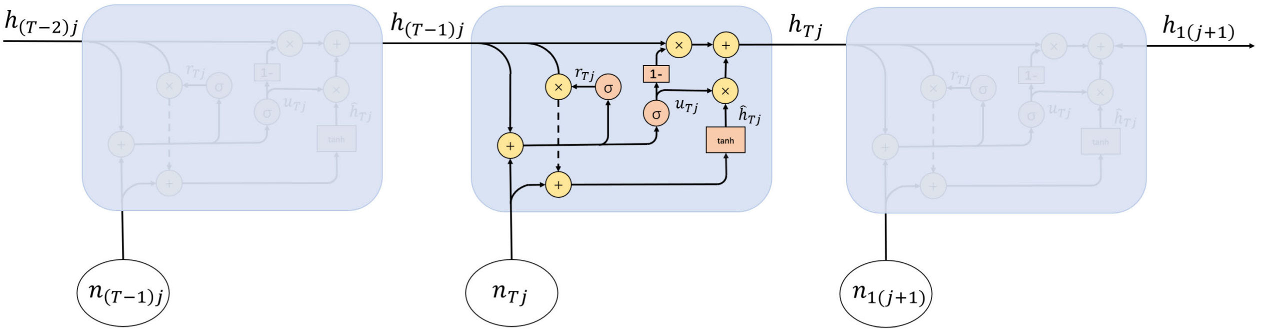


FIGURE 2. Architecture of gated recurrent unit.

calculated by:

$$u_{ij} = \begin{cases} \sigma(W_u n_{ij} + U_u h_{(i-1)j} + b_u), & i \neq 1 \\ \sigma(W_u n_{ij} + U_u h_{T(j-1)} + b_u), & i = 1 \end{cases} \tag{7}$$

where σ denotes a sigmoid activation function. W_u, U_u and b_u indicate the weight and bias parameters of the GRU model. n_{ij} is the node vector of graph structure G.

The hidden state $\hat{h}_{ij}$ is presented as follows, which is trained to capture the relationship over different time states through reset gate r_{ij}:

$$\hat{h}_{ij} = \begin{cases} tanh(W_h n_{ij} + U_h(r_{ij} \circ h_{(i-1)j}) + b_h), & i \neq 1 \\ tanh(W_h n_{ij} + U_h(r_{ij} \circ h_{T(j-1)}) + b_h), & i = 1 \end{cases} \tag{8}$$

where W_h, U_h and b_h indicate the weight and bias parameters of the GRU model, and $\circ$ is an element-wise multiplication. Moreover, the reset gate r_{ij} in (8) controls the degree to forget the previous features based on the previous state $h_{(i-1)j}/h_{T(j-1)}$ and the current node vector n_{ij}, which is computed as follows:

$$r_{ij} = \begin{cases} \sigma(W_r n_{ij} + U_r h_{(i-1)j} + b_r), & i \neq 1 \\ \sigma(W_r n_{ij} + U_r h_{T(j-1)} + b_r), & i = 1 \end{cases} \tag{9}$$

where the calculation of r_{ij} is similar to that of the update gate u_{ij}. In addition, σ, W_r, U_r, and b_r respectively denote a sigmoid activation function and the parameters of the GRU model.

In the end of the sequence G, we regard the memory cell h_{TK} as the final feature vectors. As illustrated in Figure 1, the updated features after the GCN and GRU modules are forwarded into an SoftMax classifier.

$$P = Softmax(W_p h_{TK} + b_p) \tag{10}$$

where W_p and b_p are the weights and bias of a fully connected layer that compresses the input dimension of $2048 \times k$ into only seven dimensions for subsequent emotion classification. The loss function for optimization calculates the cross-entropy loss over all the nodes as follows:

$$L = -\sum_{i=0}^{6} Y_i \ln P_i \tag{11}$$

where Y is the emotion label indicator matrix.

IV. EXPERIMENTS

A. DATASETS

Our method has been evaluated on two benchmark datasets, namely, CAER and AFEW. These datasets not only have multi-angle and natural facial expressions but also retain the surrounding context around the human face. For our experiments, the surrounding context of videos is particularly important. CAER has collected the 13,201 video clips from 79 TV shows, which are manually labeled as seven basic categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. In addition to the video dataset, Lee *et al.* [14] extracted frames from the video clips to set up the static image dataset named CAER-S, which contains about 70K static images. These two datasets are randomly split into a training set (70%), a validation set (10%), and a testing set (20%). To better validate the effect of spatial-temporal contextual features for emotion recognition, we compared our method with the baseline of CAER and CAER-S. In addition, AFEW contains about 1809 video clips from TV shows or movies. These video clips have been divided into a training set (773), a validation set (383), and a testing set (593). All video clips are labeled with the same seven basic categories as those in the CAER. We also evaluate our model on the AFEW.

B. CHOOSING HYPERPARAMETERS

In the pre-training stage, the bottom-up attention model is trained with ResNet-101 as the backbone on the Visual Genomes dataset [38], which follows the same settings as [17], [39]. We train our model from scratch using RAdam optimizer [40] with learning rates initialized as 1×10^{-3} and descended the learning rate with 0.1 every 8 epochs. In addition, we use a mini-batch size of 32. Considering that CAER datasets have various video clip lengths, we randomly extract 16 consecutive video frames from each video clips at a sample rate of 10 frames per second. We further perform data augmentation operations to horizontally flip video frames.

In our model, the number (T, K) of video frames and regions are particularly significant parameters. Different parameters choices will affect the performance of our model. To achieve the best recognition performance, we carry out further experiments with different settings of video frames T

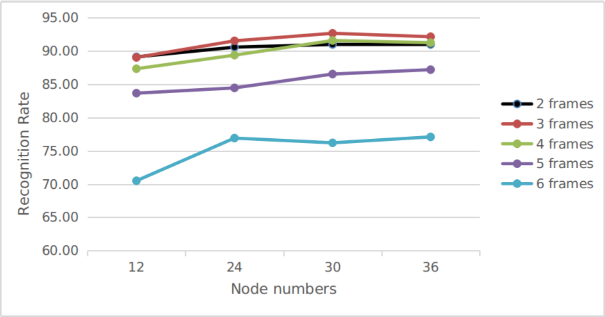


FIGURE 3. The recognition accuracy of different video frames and node numbers.

TABLE 1. Ablation studies on CAER dataset. Results are reported in terms of recognition accuracy.

Methods	Accuracy(%)
Average Pooling	36.19
GCN	68.23
GRU	71.37
1GCN+GRU	85.26
2GCN+GRU	86.06
3GCN+GRU	86.56
4GCN+GRU	86.73
5GCN+GRU	86.35

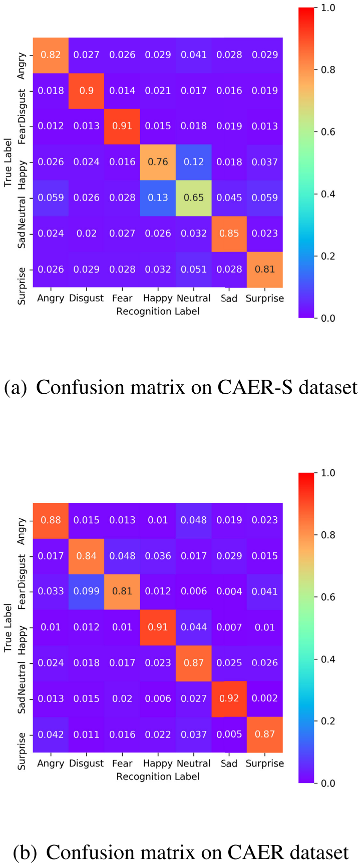
and numbers of region K. In these experiments, we select the video frames T from [2, 3, 4, 5, 6] and the number of regions K from [12, 24, 30, 36]. We analyze the emotion recognition of these experiments in the CAER dataset. As shown in the Figure 3, the horizontal axis denotes the numbers of the regions from video frames and the vertical axis denotes the accuracy of emotion recognition. In addition, we show the curves of different video frames in the different colors. These data indicate that the emotion recognition accuracy is the highest when the video frame number T is 3 and then the node number K is 30. Therefore, we set the video frame number T and node number K to 3 and 30 in our experiments, respectively.

C. ABLATION STUDIES

To evaluate our proposed methods quantitatively, we perform ablation studies to analyze each component in our framework. We propose a basic baseline model (noted as ''Average Pooling'') without any GCN and GRU layers. In the baseline model, we apply an average-pooling layer after the graph structure transition and then input the features to the emotion classifier. The emotion classifier is the same as the one used in GRERN. In Table 1, the Average Pooling model achieves 36.19% accuracy of emotion recognition. To demonstrate the capability of the GCN layer to reason the relationships between node vectors in our model, we adopt one GCN layer on the graph features to extract the emotional relationships in a similar way to the baseline model. This model is denoted as ''GCN''. The role of GRU is also validated by establishing a model marked as ''GRU''. This model adopts one GRU layer on the initial graph features without any GCN layers to capture the global emotion features. In summary, the GRU and the GRU model both generate effective emotion features and improve the emotion recognition accuracy.

In addition, we combine a GRU layer with different GCN layers to further study the performance of our proposed method. We mark these models as ''1GCN+GRU'', ..., and ''5GCN+GRU'', which respectively integrates a GRU layer with [1, 2, 3, 4, 5] GCN layers. Table 1 shows that the emotion recognition performance of our model is gradually improved by applying multiple GCN layers before the GRU layer. These results illustrate that the GCN module can extract the enhanced emotion features by learning the relationships between node vectors and that the GRU module can capture the discriminative spatiotemporal information by maintaining long-term information and removing the redundant features. The emotion recognition accuracy becomes the best when four GCN layers are added into the model by maximizing utilizing the spatiotemporal contextual information. The best recognition accuracy can be further improved by about 1% in comparison with 1GCN+GRU. Finally, we choose 4GCN+GRU as GRERN to compare with the SOTA methods.

Figure 4 demonstrates the confusion matrix of GRERN on the CAER-S and CAER datasets to analyze the recognition performance of our proposed method in each emotion



(a) Confusion matrix on CAER-S dataset

(b) Confusion matrix on CAER dataset

FIGURE 4. Confusion matrix of GRERN on the CAER-S and CAER datasets.

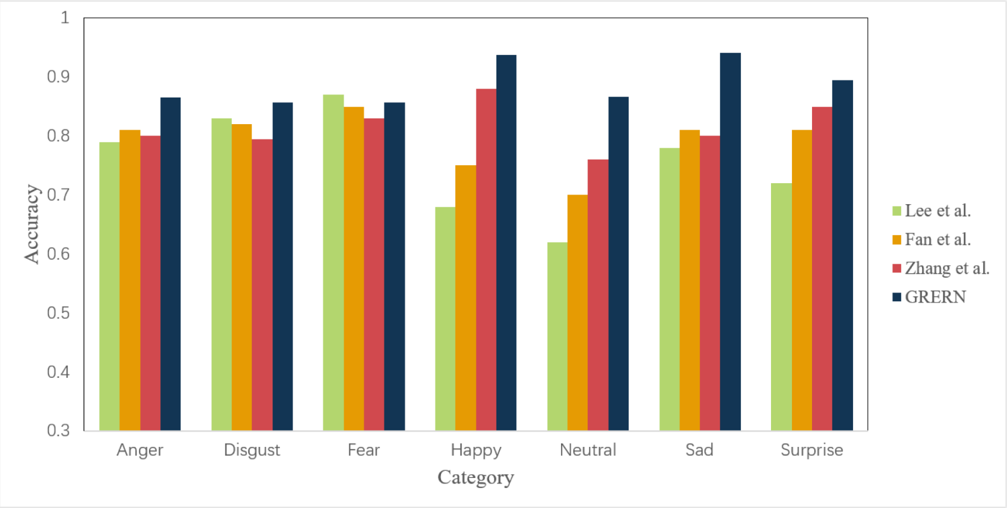


FIGURE 5. Each category of GRERN compared with baseline methods in the CAER benchmark.

category. The Neutral category has the lowest accuracy in the seven categories. As shown in the Figure 4, a large number of Neutral categories are misclassified as Happy categories in the CAER-S dataset. Moreover, when we perform GRERN on the CAER dataset, Happy and Neutral categories can be clearly distinguished. However, GRERN performs poorly in recognizing the Disgust and Fear categories, which may be attributed to the similar emotion features of the two categories in dynamic movement.

D. RESULTS OF EXPERIMENTS

To demonstrate the strength of our model in extracting the spatiotemporal information, we compare our model with several SOTA methods in the CAER and AFEW datasets. These methods are shown as follows:

1) Lee *et al.* [14] built a two-branch encoding framework. One branch achieved facial expression encoding to extract the emotional features of the human face. The other branch implemented the contextual information encoding to extract the surrounding context information around the human face. In the end, the model utilized an adaptive fusion module to fuse the extracted emotional features after the two branches.
2) Zhang *et al.* [20] established a network by integrating GCN and CNN. First, the network utilized the Region Proposal Network to convert images into node vectors. Then, those node vectors were fed into GCN to extract the emotion relationships. The body features of images were captured with CNNs. Finally, the features from GCN and CNN were connected to predict the emotion category.
3) Fan *et al.* [21] developed an architecture based on deeply supervised CNN. The model extracted multi-level and multi-scale human face features through different convolutional layers. The final features from each convolutional layer were used to predict the emotion label.

We evaluated our model and above SOTA methods on the CAER and CAER-S benchmark. However, Zhang *et al.* [20] and Fan *et al.* [21] did not provide open-sourced implementations. We have reproduced Zhang *et al.* [20] and Fan *et al.* [21] on the CAER and CAER-S datasets and compared the obtained results on the same datasets.

To quantitatively evaluate the importance of temporal information for emotion recognition, we first conduct some experiments on the CAER-S dataset. CAER-S is an image subset of CAER, which only contains the spatial features. In Table 2, the top four rows shown the emotion recognition accuracy of SOTA methods on the CAER-S dataset, and the bottom four rows illustrated the performance of SOTA methods on the CAER dataset. As shown in Table 2, the recognition performance of these methods trained on the CAER dataset is better than that of the methods trained on the CAER-S dataset. Therefore, the results demonstrate that the temporal information is beneficial for the emotion recognition.

On the CAER dataset, GRERN has achieved the best performance with an emotion recognition accuracy of 86.73%. Compared with the results of Lee *et al.* [14] and Fan *et al.* [21], the emotion recognition performance of our GRERN is improved by 9.69% and 6.01% respectively in terms of accuracy. These results indicate that our model extracts better spatiotemporal emotion features through GCN and GRU than other methods. Moreover, Zhang *et al.* [20] outperformed the method proposed in [21] with a recognition accuracy of 81.66% through combining GCN and CNN. However, GRERN combined the GRU and GCN can highly improve the performance for emotion recognition by further refining the relationships of the node vectors.

To further analyze the effectiveness of combining the GCN and GRU for emotion recognition on the CAER dataset,

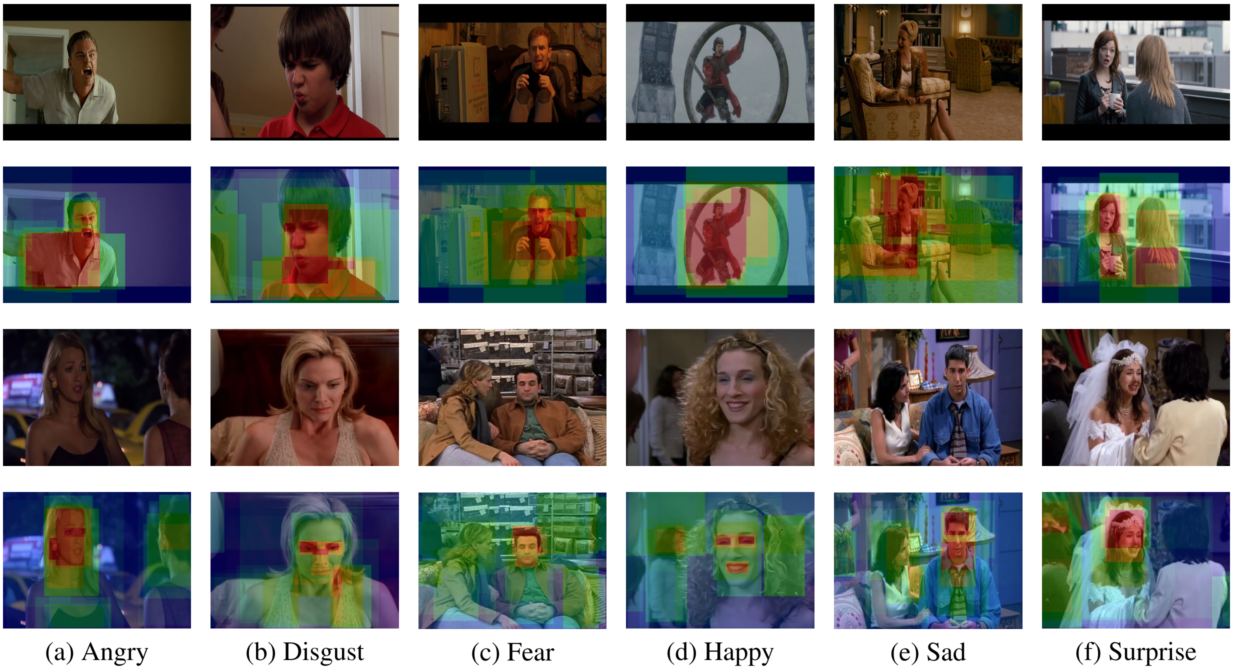


FIGURE 6. Visualization of heatmaps with the six basic emotions.

TABLE 2. Comparisons of the emotion recognition accuracy of SOTA methods on the CAER datasets.

Methods	Dataset	Accuracy(%)
Lee *et al.* [14]	CAER-S	73.51
Fan *et al.* [21]	CAER-S	75.19
Zhang *et al.* [20]	CAER-S	76.73
GRERN	CAER-S	81.31
Lee *et al.* [14]	CAER	77.04
Fan *et al.* [21]	CAER	80.72
Zhang *et al.* [20]	CAER	81.66
GRERN	CAER	86.73

TABLE 3. Comparisons of the emotion recognition accuracy of SOTA methods on the AFEW datasets.

Methods	Training Dataset	Accuracy(%)
Lee *et al.* [14]	AFEW	43.12
Fan *et al.* [21]	RAF-DB+AFEW	57.43
Zhang *et al.* [20]	AFEW	46.08
GRERN	AFEW	52.26
GRERN	CAER+AFEW	58.85

we visualize the recognition rate of each emotion category in all methods. As shown in Figure 5, the performance improvement of our model is mainly achieved by improving the recognition rates of the Happy, Sad, and Neutral categories. This result means that GRERN can better capture the dynamic changes of these categories than the above state-of-the-art methods. However, GRERN does not perform well in the recognition of the Fear category compared with the above methods. Table 4 shows the number of data for each category. The number of Fear categories in the CAER dataset is the least. Therefore, the bottleneck of GRERN for recognizing the Fear category may be caused by the lack of training data.

TABLE 4. Amount of video clips in each category on CAER dataset.

Category	CAER	CAER-S	Proportion(%)
Angry	1,628	139,681	12.33
Disgust	719	59,630	5.44
Fear	514	46,441	3.89
Happy	2,726	219,377	20.64
Neutral	4,579	377,276	34.69
Sad	1,473	138,599	11.16
Surprise	1,562	126,873	11.83
Total	13,201	1,107,877	100

We conduct more experiments on the AFEW dataset. The diffrent methods are compared on the AFEW dataset. As shown in the Table 3, GRERN is robust in context-aware emotion recognition. However, the model of Fan *et al.* [21] performs better in emotion recognition when compared with GRERN only trained with the AFEW dataset. Their model has been pre-trained on the Real-world Affective Face Database (RAF-DB) [41] which includes large facial expression samples. To verify the robustness of GRERN to context-aware emotion recognition, we pretrain GRERN on the CAER dataset and then fine-tune our model on the AFEW dataset. Finally, the performance of GRERN has been highly improved by 1.42% compared with the model of Fan *et al.* [21].

For visualizing the correlation between final features and salient regions that include the facial expression and discriminative contextual information, we calculate the similarity scores between the node vectors $G = \{N_1, \ldots, N_T\}$ and the final features h_{TK} generated in the GRU through inner product operation. Then, we color these regions with different weights according to their score ranking. Figure 6 shows

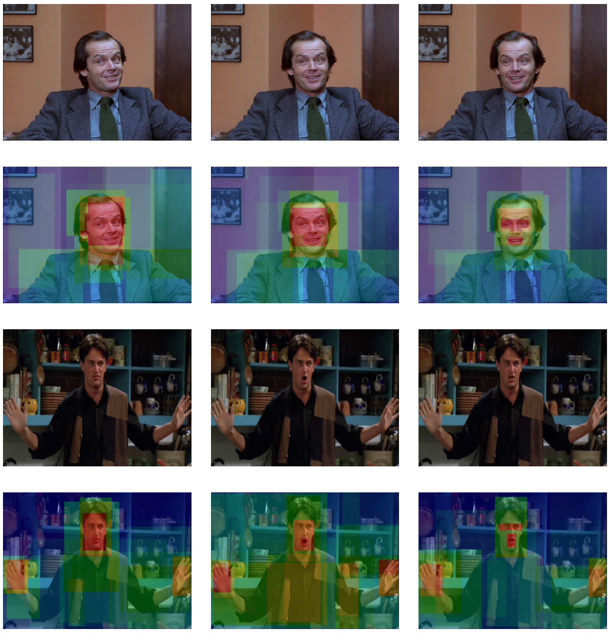


FIGURE 7. Attention visualization of video frames in CAER and AFEW datasets.

the heatmaps of each category in the AFEW and CAER datasets. The first and third rows are the each categories of original video frames in the AFEW and CAER datasets. The next rows are the corresponding heatmaps. As shown in Figure 6, GRERN can recognize the correct emotion categories through facial expressions and surrounding information even in complex scenarios. In addition, we visualize the heatmaps on multiple frames in a single video in the AFEW and CAER datasets to analyze the effectiveness of our methods in extracting temporal features. In Figure 7, the top two rows are the input video frames and heatmaps in the AFEW dataset, and the bottom two rows are the input video frames and heatmaps in the CAER dataset. As shown in Figure 7, GRERN can capture the changes in faces and gestures in time dimension to extract temporal video information. Thus, GRERN captures not only the facial expressions, but also the changes in body movements or gestures.

V. CONCLUSION

We propose the novel method GRERN that combines GCN with GRU to use the relationship between salient regions for emotion recognition. The GCN enriches the node features and constructs the connections between regions. The GRU in GRERN enables the network to remove redundant features of the graph and retain significant parts. Extensive experiments under the CAER and AFEW datasets show that GRERN outperforms state-of-the-art methods for context-aware emotion recognition. Compared with state-of-the-art methods, GRERN can better capture the spatiotemporal emotion features in video clips containing complex context information. In the future, we will conduct emotion recognition experiments on blind people or people with facial deformities.

REFERENCES

[1] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, pp. 799–823, Jan. 2015.

[2] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[3] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Muller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[4] W.-L. Zheng and B.-L. Lu, "Personalizing eeg-based affective models with transfer learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2732–2738.

[5] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[6] S. Murali, F. Rincon, and D. Atienza, "A wearable device for physical and emotional health monitoring," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2015, pp. 121–124.

[7] M. Magno, M. Pritz, P. Mayer, and L. Benini, "DeepEmote: Towards multi-layer neural networks in a low power wearable multi-sensors bracelet," in *Proc. 7th IEEE Int. Workshop Adv. Sensors Interfaces (IWASI)*, Jun. 2017, pp. 32–37.

[8] I. C. Jeong, D. Bychkov, and P. C. Searson, "Wearable devices for precision medicine and health state monitoring," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1242–1258, May 2019.

[9] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG pattern analysis for emotion detection," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 102–115, Jan. 2012.

[10] M. Merone, P. Soda, M. Sansone, and C. Sansone, "ECG databases for biometric systems: A systematic review," *Expert Syst. Appl.*, vol. 67, pp. 189–202, Jan. 2017.

[11] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[12] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[13] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[14] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10143–10152.

[15] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 435–442.

[16] C. Chen, Z. Wu, and Y.-G. Jiang, "Emotion in context: Deep semantic feature fusion for video emotion recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 127–131.

[17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[19] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," Austral. Nat. Univ., Canberra, ACT, Australia, Tech. Rep. TR-CS-11, 2011, p. 1, vol. 2.

[20] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 151–156.

[21] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Video-based emotion recognition using deeply-supervised neural networks," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 584–588.

[22] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[23] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3304–3311.

[24] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, ''Local binary patterns and its application to facial image analysis: A survey,'' *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[25] Y. Li, J. Zeng, S. Shan, and X. Chen, ''Occlusion aware facial expression recognition using CNN with attention mechanism,'' *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[26] 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.

[27] 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.

[28] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, ''EmoNets: Multimodal deep learning approaches for emotion recognition in video,'' *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, Jun. 2016.

[29] F. Zhang, T. Zhang, Q. Mao, and C. Xu, ''Joint pose and expression modeling for facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[30] T. N. Kipf and M. Welling, ''Semi-supervised classification with graph convolutional networks,'' 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[31] S. Yan, Y. Xiong, and D. Lin, ''Spatial temporal graph convolutional networks for skeleton-based action recognition,'' 2018, *arXiv:1801.07455*. [Online]. Available: http://arxiv.org/abs/1801.07455

[32] L. Yao, C. Mao, and Y. Luo, ''Graph convolutional networks for text classification,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7370–7377.

[33] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, ''Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations,'' in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 10–16.

[34] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, ''DialogueGCN: A graph convolutional neural network for emotion recognition in conversation,'' 2019, *arXiv:1908.11540*. [Online]. Available: http://arxiv.org/abs/1908.11540

[35] Y. Lai, L. Zhang, D. Han, R. Zhou, and G. Wang, ''Fine-grained emotion classification of Chinese microblogs based on graph convolution networks,'' 2019, *arXiv:1912.02545*. [Online]. Available: http://arxiv.org/abs/1912.02545

[36] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, ''ARBEE: Towards automated recognition of bodily expression of emotion in the wild,'' *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 1–25, Jan. 2020.

[37] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, ''Visual genome: Connecting language and vision using crowdsourced dense image annotations,'' *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[39] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, ''Visual semantic reasoning for image-text matching,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4654–4662.

[40] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, ''On the variance of the adaptive learning rate and beyond,'' 2019, *arXiv:1908.03265*. [Online]. Available: http://arxiv.org/abs/1908.03265

[41] S. Li and W. Deng, ''Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,'' *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

QINQUAN GAO received the B.S. degree in automation and the M.S. degree in systems engineering from Xiamen University, China, in 2008 and 2010, respectively, and the Ph.D. degree from Imperial College London, in 2014. He is currently an Associate Professor with Fuzhou University, working on model compressing, machine learning, biomedical image processing, and computer vision.

HANXIN ZENG is currently pursuing the master's degree with the College of Physics and Information Engineering, Fuzhou University. His main research interests include emotion recognition and image super-resolution.

GEN LI received the B.S. degree in computer science and technology from Southwest University for Nationalities, Chengdu, China, in 2006, and the Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2017. Since 2017, he has been a Chief Research Scientist with Imperial Vision Technology, Fuzhou, China. His current research interests include deep learning, computer vision, image analysis, and pattern recognition.

TONG TONG received the Ph.D. degree from Imperial College London, in 2015. He was a Research Fellow with the MGH/Harvard Medical School, in 2016. He is currently a Full Professor with the College of Physics and Information Engineering, Fuzhou University. His research interests include machine learning, medical image analysis, and computer aided diagnosis.