Human Emotion Recognition with Relational Region-Level Analysis

Weixin Li, Xuan Dong*, and Yunhong Wang, *Fellow Member, IEEE*

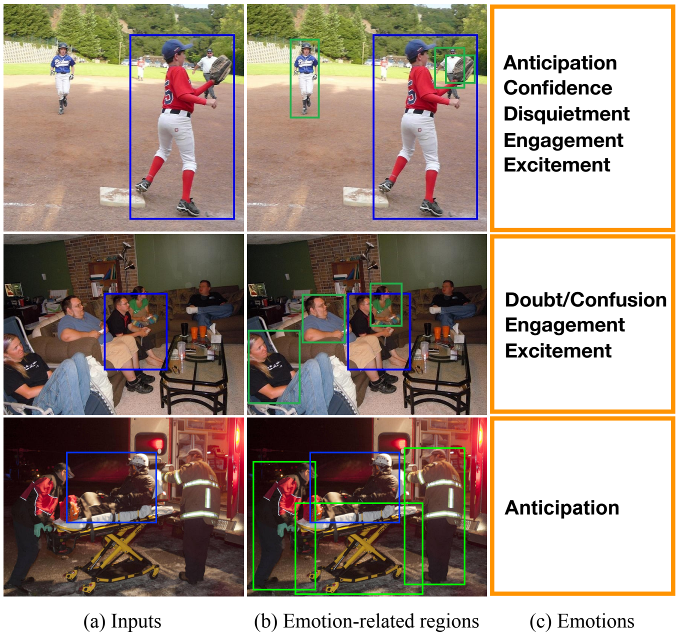
Abstract—Recognizing the emotional state of a person within the image in real-world scenarios is a key problem in affective computing and has various promising applications. Local regions in the image, including different objects in the background scene and parts within the foreground body, usually have different contributions to emotion perception of the target person. This, however, has not been well exploited in most existing methods. In this paper, we propose to make relational region-level analysis to account for the different contributions of different regions to emotion recognition. For the background scene, we propose a Body-Object Attention (BOA) module to estimate the contributions of background objects to emotion recognition given the target foreground body. Within the foreground body, we propose a Body Part Attention (BPA) module to recalibrate the channel-wise body feature responses to attend on body parts that are more important. Moreover, we propose to model the emotion label dependency in real-world images, considering both the semantic meanings of these labels and their co-occurrence patterns. We evaluate the proposed method on the EMOTIC and CAER-S datasets, and experimental results show the superiority of our method compared with the state-of-the-art algorithms.

Index Terms—Human Emotion Recognition, Real-World Scenarios, Relational Region-Level Analysis, Body-Object Attention, Body Part Attention, Emotion Label Dependency Modeling

1 Introduction

We study the problem of human emotion recognition in real-world scenarios. As shown by example images in Fig. 1, we aim at inferring the emotional state of one specific person in the image (marked by the blue rectangle) in non-controlled environments. The need of human emotion recognition has been widely acknowledged in both industry and academia [1]. Recent years also witness its extensive applications in human-computer interaction [2], [3], healthcare [4], [5], digital entertainment [6], [7], etc.

Traditional human emotion recognition methods only have faces [8], [9], [10], [11], [12] or/and bodies [13], [14] as inputs, while ignoring the fact that in practice, faces and bodies never appear in isolation. As supported by Psychology studies [15], a plenty of cues in the background scene can contribute to the human emotion recognition in real-world scenarios. Some recent works provide a good start to infer human emotions by taking the background scene into consideration [16], [17]. Kosti et al. [16] treat both of the foreground body and background scene as a whole region respectively, and use two convolutional neural network (CNN) streams with identical structures to extract their features for emotion recognition. Lee et al. [17] treat the face of the target person as the foreground to exploit human facial expression, and propose to model cues in the background by hiding the face, for emotion recognition. However, the common situation where human faces are invisible is not considered in their method, and other body parts except faces are modeled in the same way as other background cues. In short, there still exists much space for further improvement.



(a) Inputs (b) Emotion-related regions (c) Emotions

Fig. 1: Examples of emotion recognition in real-world scenarios, which aims at inferring the emotional state (c) of the target person (marked in blue in the input image (a)). (b) shows that among all regions, some regions (marked in green) are more related and important for emotion recognition.

We notice that, for human emotion recognition in real-world scenarios, different local regions in the image usually contribute differently for understanding the target person's emotional state.

- * Corresponding author.
- Weixin Li is with the School of Computer Science and Engineering, Beihang University, Beijing, 100191, China. E-mail: weixinli@buaa.edu.cn
- Xuan Dong is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China. E-mail: dongxuan8811@bupt.edu.cn
- Yunhong Wang is with the School of Computer Science and Engineering, Beihang University, Beijing, 100191, China. E-mail: yhwang@buaa.edu.cn

Manuscript received XXX, 2019.

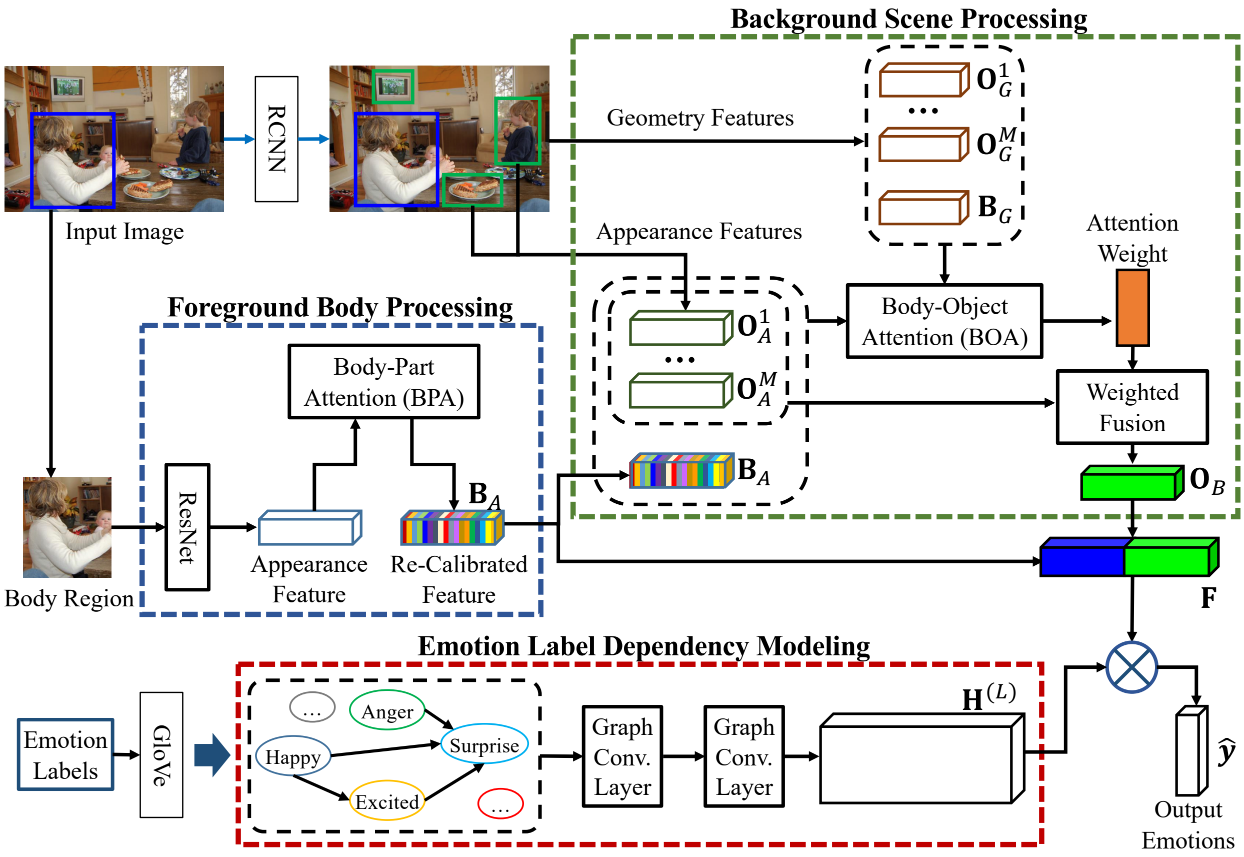


Fig. 2: The overall structure of our model, which makes relational region-level analysis for emotion recognition in real-world scenarios. The Body-Object Attention (BOA) and Body-Part Attention (BPA) are proposed to attend on object regions and body parts that are more important for emotion recognition in the background and foreground, respectively. The Emotion Label Dependency modeling (ELD) module exploits the correlations between different emotion labels for more accurate emotion recognition. (Better viewed in color.)

In the background scene, some regions (as shown by the green rectangles in Fig. 1) are more helpful for our perception of the target person, including objects that the person is interacting with, and other people nearby. Within the target foreground body region, different parts, e.g. faces, arms, legs, also contribute to the emotion recognition differently in specific scenarios.

In addition, real-world images generally contain multiple emotion labels for the target person, and strong label co-occurrence usually exists. For instance, a person in the image can be both *happy* and *excited* (e.g. when he/she is playing games), while the emotions *happy* and *sad* almost never co-occur on the same person. Moreover, positive and negative emotions usually do not appear simultaneously. Therefore, both the co-occurrence patterns of different emotions and their semantic meanings can be useful for emotion recognition.

Motivated by the above insights, we propose an end-to-end deep neural network which makes relational region-level analysis for emotion recognition in real-world scenarios. The overall structure of our model is shown in Fig. 2. For the background scene, we propose a Body-Object Attention (BOA) module to estimate the contributions of different background objects given the target foreground body, based on their appearance and geometry features. Features of these background object regions are then reweighted and fused based on the estimated attention weights. For the foreground body region, we propose a Body Part Attention (BPA) module. Based on our observation that activations of channels in the human emotion recognition network are closely related to body parts, the BPA module recalibrates the channel-wise feature responses to focus on parts that are more helpful for emotion recognition. To exploit the dependency across emotion labels for more accurate emotion recognition, we propose an Emotion Label Dependency modeling (ELD) module based on the Graph Convolutional Network [18]. Different emotion labels are modeled as nodes in the graph, and both the semantic meanings of emotion labels themselves and their co-occurrence patterns are considered in the proposed module. The output from the ELD module is further integrated with features from the foreground body and background regions for the final emotion prediction.

Our experiments are conducted on the EMOTIC dataset [16], [19] and the CAER-S dataset [17]. Both datasets are for emotion recognition in real-world scenarios. Experimental results show that our method could largely outperform the state-of-the-art algorithms.

In general, this paper makes the following contributions:

- We propose to solve the problem of human emotion recognition in real-world scenarios using relational region-level analysis.
- We propose a Body-Object Attention (BOA) module for estimating the contributions of objects in the background scene to emotion perception of the foreground target person.
- We design a Body Part Attention (BPA) module for refining features of the target foreground body, so that more

important body parts can be focused.
- We design an Emotion Label Dependency modeling (ELD) module based on Graph Convolutional Network for modeling emotion label dependency using both their semantic meanings and co-occurrence patterns.

The rest of this paper is organized as follows. We firstly review the related work in Section 2. Then we present the proposed method in Section 3. We report our experiment results and comparisons with other state-of-the-art methods in Section 4 and conclude in Section 5.

2 Related Works

2.1 Face and Body Based Human Emotion Recognition

Most existing studies for human emotion recognition focus on faces, assuming that emotional states of people can be inferred based on their facial expressions. Over the past years, facial expression recognition has been extensively studied [8], [9], [20], [21], [22], [23]. Early works mainly use face images captured under controlled lab environments [24], which include limited variations of head poses, illumination conditions, etc. Recent studies explore facial expression recognition in the wild [22], [23], and the expressions are also spontaneous and with diverse poses. As for the algorithms used for facial expression recognition, traditional methods mostly use hand-crafted appearance and geometry features extracted either from the whole face or specific local face regions, e.g. Scale-Invariant Feature Transform (SIFT) [9], Local Binary Pattern (LBP) [9], [24], Pyramids of Histograms Of Gradients (PHOG) [20], which are then fed into supervised classifiers e.g. Support Vector Machine (SVM) [25], random forest [21], etc., to infer human emotions. Recent works are mostly deep learning based ones, which use Convolutional Neural Networks (CNNs) to jointly extract facial features and recognize emotions [8], [11], [22], [23], and achieve satisfactory performance.

Since body gestures play an important role in conveying emotions as well [26], some other methods utilize hand, shoulder, body poses, etc., for emotion recognition. Karpouzis et al. [27] extract emotion-related features through hand movements. Nicolaou et al. [28] fuse shoulder gesture cues with those from facial expressions to predict emotions. Schindler et al. [29] propose a neural model to recognize emotions based on the body pose. Yang and Narayanan [13] model body gesture dynamics to recognize the emotional states of persons in a dyadic interaction. Deep learning for body emotion recognition has also been explored recently [30], [31], [32]. Barros et al. [30] propose a Multichannel CNN for emotion recognition with face and upper body. In [32], Nguyen et al. propose a novel feature-level fusion approach based on multimodal compact bilinear pooling for fusing multimodal emotion cues, including facial expressions, poses, and body movements.

The limitations of face-based and body-based emotion recognition algorithms are that they only focus on analyzing face and body regions of the target person. However, in real-world scenarios, plenty of cues from the background scene of the image can be utilized for emotion recognition, which are ignored by these algorithms. Moreover, the face and foreground body may have occlusions or even be invisible, which can hardly be handled by these algorithms.

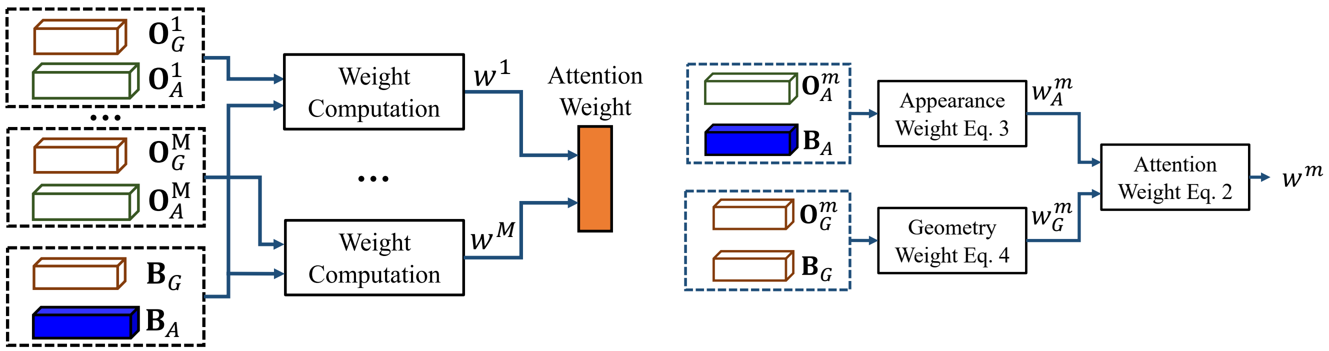
2.2 Human Emotion Recognition in Real-World Scenarios

In real-world scenarios, an individual's face and body are usually accompanied with the background scene, which can contribute substantially to his/her emotion perception [15], [33]. Recently, Kosti et al. [16], [19] make use of both foreground body and background scene to recognize emotions of the target person in the image. They also present a dataset, named EMOTIC, with images containing people in contexts under natural environments. However, they treat the whole background scene and target foreground body as single regions to extract features for emotion recognition. But, different regions/parts have different levels of importance for emotion recognition and their method thus can not make full use of the important cues. Lee et al. [17] propose the Context-Aware Emotion Recognition Networks (CAER-Net) for human emotion recognition in real-world scenarios. They exploit the scene contexts by hiding the human faces in the image, and model their contributions in a joint and boosting manner together with those of the human face areas. Moreover, they build a dataset called Context-Aware Emotion Recognition (CAER), which contains a large number of TV show video clips with labeled emotion categories. Their proposed method, however, has not modeled local regions' contributions meticulously, and can hardly deal with occluded/invisible faces which commonly happens in real-world scenarios. Zhang et al. [34] construct an affective graph to utilize contexts for emotion recognition based on Graph Convolution Network. However, the background cues are only used to enrich the foreground body features, while not considered thoroughly for human emotion recognition. The geometry features of foreground and background cues are not exploited as well. Mittal et al. [35] propose to perform emotion recognition based on multiple modalities including faces and gaits of the target person, as well as the background scene, while the analysis is not detailed enough to model their specific contributions. Some body part cues, e.g. the body gesture are also ignored in the emotion perception.

To sum up, existing methods for human emotion recognition in real-world scenarios have not solved well in terms of the utilization of cues provided by local regions in the foreground body and background scene.

2.3 Attention Mechanisms in CNNs

The attention mechanism has been widely and successfully used in convolutional neural networks for various tasks, e.g. machine translation [36], image caption generation [37], object detection [38], scene segmentation [39], etc. Vaswani et al. [36] propose the first sequence transduction model Transformer, which is designed with multi-headed self-attention. Hu et al. [38] model relations between objects for object detection based on the attention mechanism. Fu et al. [39] propose to adaptively integrate local features with their global dependencies via a Dual Attention Network (DANet). Lee et al. [17] use the attention mechanism to seek salient context information in the background scene by hiding the face region. In this paper, we propose to model the contributions of the background local regions and different body parts to the emotion recognition of the foreground person in real-world scenarios, respectively, through relational region level analysis based on the attention mechanism.



(a) The BOA module.

(b) The weight computation operation.

Fig. 3: Illustration of the BOA module and details of its weight computation operation.

2.4 Label Dependency Modeling

Modeling label dependency is a critical problem in multi-label classification. In the field of computer vision, since images usually correspond to multiple labels, e.g. a set of objects commonly co-exist in the image, multi-label image classification has been actively studied, where the label dependency is trying to be captured for more accurate predictions. Recently, many methods have been proposed to apply deep learning for multi-label image classification. Wang et al. utilize the recurrent neural networks (RNNs) to model higher-order label relationships [40]. The attention mechanism has also been applied to multi-label image classification [41], [42], which captures semantic and spatial correlations between labels based on attention maps. Some other methods use the graphs to model the label dependency, which represent label correlations in the graph, and learn related structures and parameters in various ways, e.g based on conditional graphical Lasso (CGL) [43], Graph Convolutional Network (GCN) [44], etc. For emotion recognition, previous studies are mostly based on the basic emotion categories which are mostly treated as mutually exclusive, so the emotion label dependency has not been well explored. However, emotions in real-world images are usually more complex than the basic emotions, and strong correlations between different emotions actually exist. The method of Ruan et al. [45] models the inner connections among labels based on an encoder-decoder framework, and treats the multi-label classification task as a sequence generation problem. But the label co-occurrence patterns are not modeled in their method.

2.5 Visual Sentiment Analysis

Visual Sentiment Analysis (VSA) is a problem that is related to human emotion recognition, which aims to predict the emotional reactions of humans towards visual stimuli e.g. images [46], [47], [48]. However, the image is not required to contain a person, and even if persons exist in the image, the sentiment of the whole image is still not always consistent with the emotional states of these persons. Local regions have been specifically processed in VSA, e.g. You et al. [47] discover the local regions relevant to the image sentiment based on the attention mechanism, which is achieved by jointly learning their corresponding weights guided by descriptive visual attribute recognition, and building sentiment classifiers. Song et al. [48] generate the attention distribution over the regions of the image with the saliency map as the prior for sentiment prediction. In this paper, instead of modeling the relevance of local regions to the sentiment of the whole image, we jointly train BPA and BOA for attending on important local regions with the target person and in the background, respectively, by estimating their contributions with respective to the emotion of the target person.

3 Method

We propose to make relational region-level analysis for human emotion recognition in real-world scenarios. The structure of our model is shown in Fig. 2.

To model the contributions of different objects in the background scene to human emotion recognition, we propose a body-object attention (BOA) module, which calculates the weights of different background objects given the foreground body based on their appearance and geometry features.

To attend on the parts of foreground body that are more related to human emotion recognition, we propose a body part attention (BPA) module to learn more representative features by recalibrating channel-wise feature responses, based on our observation that these channel activations are closely related to body parts.

We also propose to model the emotion label dependency based on the Graph Convolutional Network. Both the semantic meanings of emotion labels and their cooccurrence patterns in the training set are utilized in the proposed emotion label dependency modeling (ELD) module. The output from the module is integrated with the reweighted appearance features of the background object regions and the refined foreground body appearance features to get the final emotion recognition results.

3.1 BOA Module

The BOA module estimates contributions of different objects in the background scene to the emotion perception of the foreground target person. Appearance features of the background objects are reweighted with respect to the foreground body, and then fused for emotion recognition. Geometry features are also utilized in the module to account for the relative distances and sizes between the background objects and the foreground body.

Specifically, for the foreground body $\mathbf{B}$, we denote its geometry feature and appearance feature as $\mathbf{B}_G$ and $\mathbf{B}_A$, respectively. For each object in the background region, we denote its geometry and appearance features as $\mathbf{O}_G^m$ and $\mathbf{O}_A^m$, respectively, where $m = 1, ..., M$, and M is the total number of background objects.

As shown in Fig. 3, with respect to the target foreground body $\mathbf{B}$, we model contributions of background objects to the recognition of the target person's emotional state in the BOA module via:

$$\mathbf{O_B} = \sum_{m=1}^{M} w^m \mathbf{O}_A^m, \tag{1}$$

where $\mathbf{O_B}$ is a weighted sum of appearance features of background objects. The attention weight w^m describes the contribution of the m-th background object to the emotion recognition of the foreground body:

$$w^m = softmax(w_A^m + \log(w_G^m)), \tag{2}$$

where w_G^m and w_A^m are respectively the geometry and appearance weights.

Specifically, the appearance weight is computed as:

$$w_A^m = dot(W_1 \mathbf{O}_A^m, W_2 \mathbf{B}_A)/\sqrt{d_m}, \tag{3}$$

where the appearance features $\mathbf{O}_A^m$ and $\mathbf{B}_A$ are first linearly transformed by the matrices W_1 and W_2 respectively so that their similarity can be measured in the projected subspace, and d_m is the dimension of feature after the linear transformation .

The geometry weight is computed as:

$$w_G^m = W_G \cdot E(\mathbf{O}_G^m, \mathbf{B}_G), \tag{4}$$

where E is the embedding of a relative geometry feature $[\log(\frac{|x_m - x_B|}{w_B}), \log(\frac{|y_m - y_B|}{h_B}), \log(\frac{w_m}{w_B}), \log(\frac{h_m}{h_B})]^T$, which describes the distances between the top-left corners (x_B, y_B) and (x_m, y_m) of the bounding boxes of the foreground body and background object m, and the ratios of their widths and heights (w_B, h_B) and (w_m, h_m). The embedding is calculated using the method proposed in [36], based on cosine and sine functions of different frequencies. W_G denotes the linear transformation of the obtained embedding $E(\mathbf{O}_G^m, \mathbf{B}_G)$.

To combine the appearance features of foreground body $\mathbf{B}_A$ and the reweighted and fused appearance features of background regions $\mathbf{O_B}$, we directly use the concatenation operation:

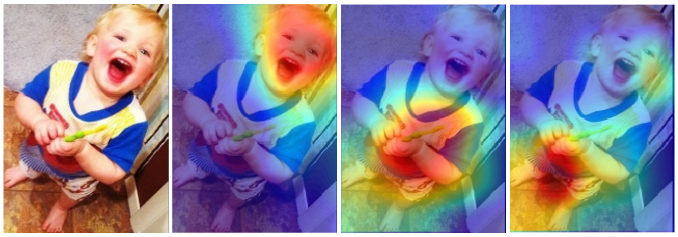
$$\mathbf{F} = concat(\mathbf{B}_A, W_3 \mathbf{O_B}), \tag{5}$$

where $\mathbf{O_B}$ is linearly transformed by W_3 before the concatenation with $\mathbf{B}_A$, and $W_3\mathbf{O_B}$ has the same feature dimension as $\mathbf{B}_A$. Since the multi-head attention, which performs single attention h times to attend to information from different representation subspaces, is proved to be useful [36], we also adopt it here. Specifically, the calculation of $\mathbf{O_B}$ is performed h times, and all the obtained results are firstly concatenated and linearly transformed (so that the obtained features still have the same dimension as $\mathbf{B}_A$) before the concatenation with $\mathbf{B}_A$.

3.2 BPA Module

For the foreground body, different parts, e.g. face, arm, leg, etc., may contribute to the perception of emotional states differently. So we propose to use the attention mechanism to guide the network to pay more attention to the emotion-related body parts, i.e. refining the foreground body feature $\mathbf{B}_A$.

Inspired by recent works showing that some filter responses in CNNs are closely related to semantic parts [49], [50], we firstly investigate whether channels in the CNN for emotion recognition activate responses for different body parts. We use the ResNet trained merely based on the foreground body for emotion recognition and visualize the heat maps of activations from different channels of the convolution layer. The heat maps of three representative channels are shown in Fig. 4. As we can see from this figure, the channel activations are closely related to specific parts, i.e. face and arms, and legs, respectively.



(a) Body region. (b) Heat maps of three channel activations.

Fig. 4: Heat maps of different channel features of ResNet trained for emotion recognition. This figure shows that the channel activations are closely related to body parts, which motivates the proposed BPA module.

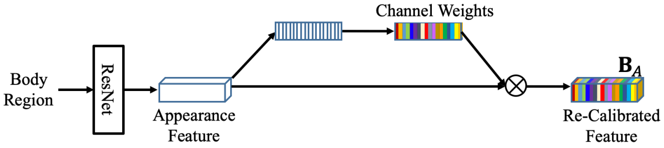


Fig. 5: Illustration of the BPA module.

Our observation is also supported by studies from other researchers showing that some detection and classification networks have channels closely associated with parts [50], [51]. Accordingly, we propose to use channel-wise attention in emotion recognition of the foreground body, to make the network focus on parts that are more helpful.

In order not to include extra training data, we adaptively recalibrate channel-wise feature responses using the squeeze-and-excitation (SE) block [52]. Specifically, as shown in Fig. 5, the output of the last convolution layer in ResNet module is firstly passed through a global pooling layer to generate a channel-wise descriptor, i.e. the feature maps are aggregated across their spatial dimensions in the squeeze operation. Per-channel modulation weights are then generated based on a self-gating mechanism in the excitation operation. These weights are further applied to the original feature maps to generate the final outputs of the BPA module, i.e. the re-calibrated body part features.

3.3 ELD Module

In the ELD module, we exploit the dependency across emotion labels explicitly for more accurate emotion recognition based on GCN [18]. Both the semantic meanings of emotion labels themselves and their co-occurrence patterns are considered in the proposed module. Specifically, for emotion labels, we extract their word embeddings $\mathbf{Z} \in \mathbb{R}^{C \times d_z}$ based on GloVe [53], where C is the number of emotion labels and d_z denotes the dimension of the embedding. The co-occurrence matrix of emotion label pairs in the training set is denoted as $\mathbf{P} \in \mathbb{R}^{C \times C}$. $\mathbf{P}_{ij}$ is the probability that the j-th emotion label appears when the i-th one exists and is calculated as the number of co-occurring pairs divided by the number of the i-th emotion label in the training set ($\mathbf{P}_{ii}$ is set as 0). We visualize $\mathbf{P}$ based on the training set of the EMOTIC dataset (where details of the dataset and the annotated emotion

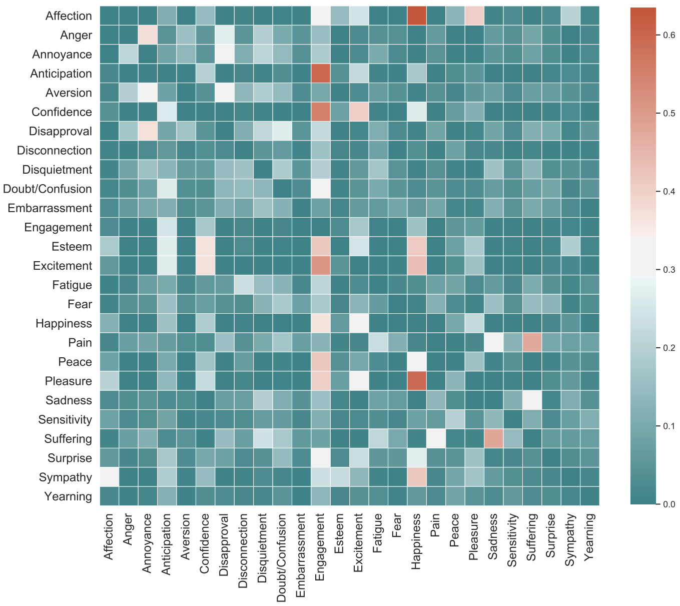


Fig. 6: Visualization of the co-occurrence matrix $\mathbf{P}$ of emotion label pairs in the training set of the EMOTIC dataset. The matrix demonstrates the existence of label dependency across different emotion labels.

categories can be found in Sec. 4) in Fig. 6, which shows the existence of label dependency across different emotion labels.

GCN primally learns node representations by encoding both node features and local graph structure [18]. Formally, given a graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ and $\mathcal{E}$ respectively denote the vertices and edges between them, and $\mathbf{A}$ is the corresponding adjacency matrix. Each node is associated with a feature vector and here we use $\mathbf{H} \in \mathbb{R}^{K \times d_v}$ to denote the feature matrix for all K nodes in the graph where d_v is the feature dimension. For one L-layer GCN, which consists of L graph convolution layers, each layer learns the embedding for each node by mixing the embeddings of its neighbors in the graph from the previous layer via:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \tag{6}$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{K \times d_v^l}$ denotes the node embeddings at the l-th layer (and we have $\mathbf{H}^{(0)} = \mathbf{H}$), $\hat{\mathbf{A}} \in \mathbb{R}^{K \times K}$ is the normalized and regularized adjacency matrix, $\mathbf{W}^{(l)} \in \mathbb{R}^{d_v^l \times d_v^{l+1}}$ is the transformation matrix to be learned, and $\sigma(\cdot)$ is the activation function which is usually set as the element-wise ReLU.

For our problem, we treat each emotion category as one node in the graph, and thus we have $\mathbf{H} = \mathbf{Z}$ and $K = C$. The adjacency matrix $\mathbf{A}$ in GCN can be calculated as the binarized matrix of $\mathbf{P}$. Specifically, $\mathbf{A}_{ij}$ is set as 1 if $\mathbf{P}_{ij} \geq t$ and 0 otherwise, where the threshold t is used to remove noisy edges. Inspired by the work in [54], which proposes a re-weighted correlation matrix to alleviate the over-smoothing problem of binary correlation matrix, we calculate the re-weighted adjacency matrix $\mathbf{A}^r$ in GCN as:

$$\mathbf{A}_{ij}^r = \begin{cases} p / \sum\limits_{j=1, i \neq j}^{C} \mathbf{A}_{ij}, & \text{if } i \neq j, \\ 1 - p, & \text{otherwise}, \end{cases} \tag{7}$$

where p determines the weights assigned to a node with respect to other correlated nodes. The output of GCN, i.e. $\mathbf{H}^{(L)}$, will be multiplied with the image representation learned in Eq. 5 to calculate the final emotion classification result $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \mathbf{H}^{(L)}\mathbf{F}. \tag{8}$$

So we have the feature dimension of $\mathbf{H}^{(L)}$, i.e. d_v^L, set as equal to the feature dimension of $\mathbf{F}$.

4 Experiments

4.1 Datasets

Our experiments are conducted on two datasets for emotion recognition in real-world scenarios, namely the EMOTIC dataset [16], [19], CAER-S dataset [55].

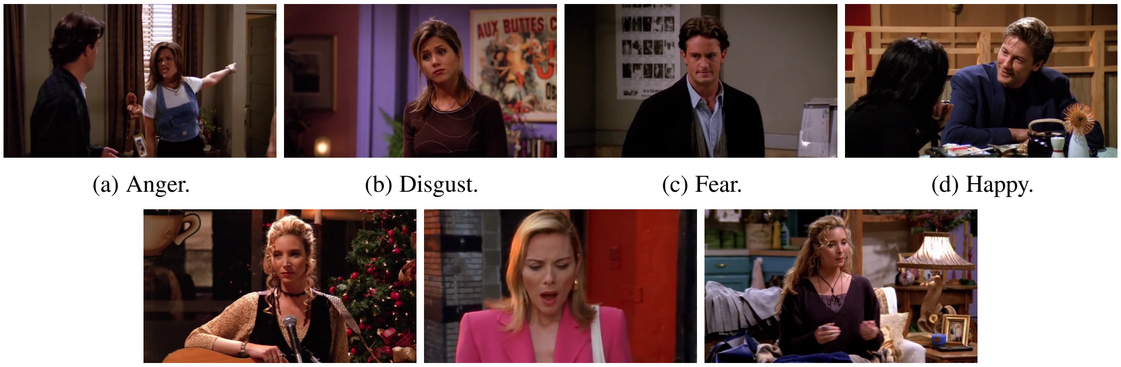
The EMOTIC dataset is for emotion recognition in real-world scenarios, which contains a total number of 23,571 images with 34,320 annotated people. Each person is labeled with 26 discrete emotion categories and continuous Valence, Arousal, and Dominance dimensions. The 26 emotion categories cover a wide range of emotional states, including *Peace*, *Affection*, *Esteem*, *Anticipation*, *Engagement*, *Confidence*, *Happiness*, *Pleasure*, *Excitement*, *Surprise*, *Sympathy*, *Doubt/Confusion*, *Disconnection*, *Fatigue*, *Embarrassment*, *Yearning*, *Disapproval*, *Aversion*, *Annoyance*, *Anger*, *Sensitivity*, *Sadness*, *Disquietment*, *Fear*, *Pain*, and *Suffering*. The detailed definitions of these emotion categories can be found in [16], [19]. The continuous Valence, Arousal, and Dominance (VAD) values measure how pleasant or unpleasant an emotion is, how likely the person is to take action under the emotional state, and the sense of control over the emotion [56], respectively. Since we focus on the emotion classification problem in this paper, only the 26 discrete categories are used in the experiments. Manually annotated body regions are provided in the EMOTIC dataset. The training, validation, and testing sets are split in the same way as [16], where numbers of samples in these three sets are respectively 70%, 10%, and 20% of the total number of the whole dataset. Some example images from the EMOTIC dataset can be found in Fig. 1.

The CAER-S dataset contains 70K static images collected from 79 TV shows, which is a subset of the CAER dataset [55]. Video clips in TV shows are processed and refined by shot boundary detector, face detector/tracking and feature clustering. Then each refined video clip is annotated with basic emotion categories including *Anger*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise*, and *Neutral*. The 70K static images in CAER-S dataset are extracted from these refined and annotated video clips, which are further randomly split into training (70%), validation (10%) and testing (20%) sets. Some example images from the CAER-S dataset for the seven basic emotion categories are shown in Fig. 7.

4.2 Implementation Details

We use the 50-layer ResNet [57] as the backbone network to extract features for the foreground body, and initialize the module using models pre-trained on the ImageNet datatset [58]. The feature dimension of the foreground body is 2,048.

Features from the last convolution layer of the 50-layer ResNet are used as the input for the body part attention module to guide the generation of channel-wise weights. The two fully connected (FC) layers used in the SE block are 32-d and 512-d, which are followed by ReLU and Sigmoid, respectively. The recalibrated features are then used as the appearance features of the target foreground body $\mathbf{B}_A$.



(a) Anger. (b) Disgust. (c) Fear. (d) Happy.

(e) Sad. (f) Surprise. (g) Neutral.

Fig. 7: Example images from the CAER-S dataset for the seven basic emotion categories.

For the background scene region, we use a pre-trained Faster R-CNN [57], [59] for object detection and feature extraction. The object types include vehicles, daily necessities, persons, animals, etc. End-to-end training can also help adaptively estimate the weights of different objects for the emotion recognition. We choose the top-M detected objects and extract C4 features for each object as its appearance features, as is done in [60], and we set $M = 50$ in this paper. In the body-object attention module, we set the head number as $h = 16$, and the fully connected layers used for linear transformations W_1, W_2, W_G, W_3, are 128-d, 128-d, 128-d, and 2048-d, respectively. The feature dimension of $\mathbf{F}$ is thus 4,096.

For the emotion label dependency modeling part, the GCN we use contains two graph convolution layers, i.e. $L = 2$, with their output dimensions set as 2,048 and 4,096 respectively. The word embeddings extracted based on GloVe pre-trained on the Wikipedia dataset have features dimension $d_z = 300$. The threshold t for calculating $\mathbf{A}_{ij}$ is set as 0.2, and p in Eq. 7 is set to be 0.2 following [54].

On the EMOTIC dataset, the loss function used for training the whole model in an end-to-end way is defined as the weighted mean square error loss following [16], [19]:

$$\mathcal{L}_{\text{EMOTIC}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\alpha_j(\hat{y}_{i,j} - y_{i,j})^2, \tag{9}$$

where N is the number of training samples, $C = 26$ is the number of emotion categories, and α_i is the weight for the j-th category. We have $\alpha_j \propto 1/\ln(\lambda + p_j)$ and $\sum_{j=1}^{C}\alpha_j = 1$, where p_j is the probability of the j-th category in the training set, and λ is a parameter which is set as $\lambda = 1.2$ in this paper.

On the CAER-S dataset, since no annotated face/body regions are provided, we follow the work in [55] and use Dlib [61] to detect faces in the image for inferring the emotional states of the target person. Moreover, due to the fact that images in the CAER-S dataset are annotated with basic emotion labels which are mutually exclusive, no co-occurrence patterns of these basic emotions exist. We thus remove the Emotion Label Dependency modeling (ELD) module when training the proposed model on the CAER-S dataset. For emotion recognition, one fully connected layer after the feature extraction modules is used for inferring the seven emotion categories. The cross-entropy loss function is used for training [55].

The proposed deep convolutional network is implemented with MindSpore. All models are optimized by stochastic gradient descent with the learning rate 0.01 and momentum 0.9 on two Nvidia 1080Ti GPUs. The batch size is set as 200 for training.

4.3 Experimental Results

Results on the EMOTIC dataset. We compare our method with the state-of-the-art models for emotion recognition in real-world scenarios, i.e. the methods of Kosti et al. [16], Zhang et al. [34], Lee et al. [17], Mittal et al. [35], and Ruan et al. [45] on the EMOTIC dataset. The methods of Kosti et al. [16] and Zhang et al. [34] are trained based on 26 discrete emotion category labels as well as the continuous VAD labels. Since only 26 discrete categories are utilized in this paper, for fair comparison, we also include the method where the same model as Kosti et al. [16] is used, but is trained with merely the discrete criterion (i.e. only based on the discrete emotion category labels), from the work of Kosti [62] in the comparison. For the method of Lee et al. [17], we re-implement the model and train it based on the discrete emotion category labels (since neither results on the EMOTIC dataset nor the codes are publicly available), and we use the body instead of face of the target person in the input. The methods of Mittal et al. [35] and Ruan et al. [45] are both trained only using the discrete emotion labels [35].

The metric used for evaluating the emotion recognition performance is the Average Precision (AP) [16], which summarizes the precision-recall curve as the weighted mean of precisions achieved at each threshold, i.e. the area under the precision-recall curve. For the method of Ruan et al. [45], since no public codes are available, we directly compare with it based on metrics used in their paper, including the label-based metrics (C-F1) and example-based metrics (O-F1), where the average is calculated over all emotion categories and all testing examples respectively.

The comparison results are shown in Table 1 and Table 2. Table 1 reports the AP scores for all 26 categories and the mean average precision (mAP). As we can see from the table, our method achieves higher AP scores compared to other methods for most emotion categories. The mAP of our method is even higher than Kosti et al. [16] and Zhang et al. [34], which are trained with extra continuous VAD labels. Compared with Kosti with discrete criterion (w/ DC) [62], Lee et al. [17] and Mittal et al. [35] which have the same setting as ours, our method has better performance for most of the 26 discrete emotion categories. Table 2 reports the C-F1 and O-F1 scores of our method and the method of Ruan et al. [45], where the results clearly show the superiority of our method.

TABLE 1: Average precision (%) of different methods for each emotion category on the EMOTIC dataset.

Category	Methods					
	Kosti et al. [16]	Zhang et al. [34]	Kosti w/ DC[1][62]	Lee et al. [17]	Mittal et al. [35]	Ours
(1) *Affection*	27.85	**46.89**	19.46	23.25	41.83	37.93
(2) *Anger*	9.49	10.87	8.10	9.71	11.41	**13.73**
(3) *Annoyance*	14.06	11.27	9.79	13.43	17.37	**20.87**
(4) *Anticipation*	58.64	62.64	52.27	54.12	**67.59**	61.08
(5) *Aversion*	7.48	5.93	5.58	8.64	**11.71**	9.61
(6) *Confidence*	78.35	72.49	60.59	72.35	65.27	**80.08**
(7) *Disapproval*	14.97	11.28	8.10	15.26	17.35	**21.54**
(8) *Disconnection*	21.32	26.91	20.79	21.53	**41.46**	28.32
(9) *Disquietment*	16.89	16.94	14.66	16.81	12.69	**22.57**
(10) *Doubt/Confusion*	29.63	18.68	28.47	27.77	31.28	**33.50**
(11) *Embarrassment*	3.18	1.94	2.58	2.29	**10.51**	4.16
(12) *Engagement*	87.53	**88.56**	81.72	83.43	84.62	88.12
(13) *Esteem*	17.73	13.33	17.54	17.84	18.79	**20.50**
(14) *Excitement*	77.16	71.89	65.20	70.68	**80.54**	80.11
(15) *Fatigue*	9.70	13.26	7.87	8.91	11.95	**17.51**
(16) *Fear*	14.14	4.21	12.11	12.36	**21.36**	15.56
(17) *Happiness*	58.26	73.26	54.55	55.79	69.51	**76.01**
(18) *Pain*	8.94	6.52	4.79	9.22	9.56	**14.56**
(19) *Peace*	21.56	**32.85**	17.69	19.03	30.72	26.76
(20) *Pleasure*	45.46	57.46	42.34	43.22	**61.89**	55.64
(21) *Sadness*	19.66	25.42	9.11	10.39	19.74	**30.80**
(22) *Sensitivity*	9.28	5.99	4.09	7.34	4.11	**9.59**
(23) *Suffering*	18.84	23.39	7.41	9.71	20.92	**30.70**
(24) *Surprise*	**18.81**	9.02	16.77	13.70	16.45	17.92
(25) *Sympathy*	14.71	17.53	10.52	16.29	**30.68**	15.26
(26) *Yearning*	8.34	**10.55**	7.64	9.59	10.53	10.11
mAP	27.38	28.42	22.68	25.10	31.53	**32.41**

[1] DC is short for discrete criterion.

TABLE 2: Label-based and example-based F1 scores of different methods on the EMOTIC dataset.

Methods	C-F1 (%)	O-F1 (%)
ResNet [45]	8.18	28.35
Kosti et al. [16]	8.31	39.91
Ruan et al. [45]	13.42	45.77
Ours	**15.10**	**48.07**

The methods of Kosti et al. [16] and Kosti with discrete criterion [62] treat all regions/parts in the background scene and foreground body equally when extracting features. However, some regions/parts are more important and helpful for emotion recognition while others may not be closely related to the emotion perception of the target foreground person. The method of Zhang et al. [34] utilizes background contexts to enrich the features of the foreground body, but contributions of different parts inside the foreground body are not considered, and geometric relations between the foreground and background regions are not modeled. The method of Lee et al. [17] encodes the background contexts by hiding the foreground region, and fuses the foreground and background features based on the attention mechanism, yet relations between local regions are not well modeled. Mittal et al. [35] use multiple modalities including faces and gaits of the target person and context information to perform emotion recognition. But still the contributions of other body parts, e.g. body gestures, are not considered. And the whole image is passes through a ResNet model with attention operations to capture the background contextual cues, where the analysis is so rough that important regions cannot play the corresponding roles. Moreover, all these methods have not taken the emotion label dependency into consideration. The method of Ruan et al. [45] captures the inner connections among labels by transforming the multi-label classification task into a sequence generation problem, based on an encoder-decoder framework. However, the analysis of the background regions only includes an attention module, so the relations of different regions are not modeled. On the contrary, our method could focus more on the regions/parts that are closely related to emotions of the target foreground person via relational region-level analysis. Both appearance and geometry features are used in the BOA module. We also model the dependency among different emotion labels. Hence better performance is achieved by our method.

Some qualitative results of our methods on the EMOTIC dataset are shown in Fig. 8. For each of the 26 discrete emotion categories, we use the validation set to decide the threshold for detection where *Precision* equals *Recall*. Then for each test sample, if we denote Q_{det} as the set of detected emotions, and Q_{gt} as the set of ground-truth emotions, the Jaccard coefficient (JC) score for this sample can thus be calculated as $|Q_{det} \cap Q_{gt}|/|Q_{det} \cup Q_{gt}|$ [16]. We randomly select some qualitative results with different JC scores, where the wrongly predicted emotion categories are marked in red. As we can see from the figure, by making relational region-level analysis and label dependency modeling, our method generally can infer the emotional states of the target person effectively.

Results on the CAER-S dataset. We also evaluate the proposed method quantitatively on the CAER-S dataset. The comparison results, i.e. the classification accuracy, of different methods are shown in Table 3 and Fig. 9. The methods of Kosti et al. [16] and Zhang et al. [34] here are also trained with the cross-entropy loss. Fine-tuned ResNet [17] is initialized using pre-trained models from ImageNet, and fine-tuned on the CAER-S dataset. All the methods use the face of the target person, along with the image, as inputs, for emotion classification.

As the results demonstrate, compared to other state-of-the-art methods, our model improves the emotion classification accuracy by more than 7%, and the accuracy for each category also shows the superiority of our model.

TABLE 3: Emotion classification accuracy of different methods on the CAER-S dataset.

Methods	Accuracy (%)
Kosti et al. [16]	74.48
Zhang et al. [34]	77.02
Lee et al. [17]	73.51
Fine-tuned ResNet [17]	68.46
Ours	**84.82**

Ablation Study. An ablation study is conducted in the experiment, which compares a number of different model variants and justifies the design choices of our method, based on the EMOTIC dataset. Specifically, we wish to evaluate the three key components in this paper, namely the BOA, BPA, and ELD modules. Table 4 shows the performance of different model variants.

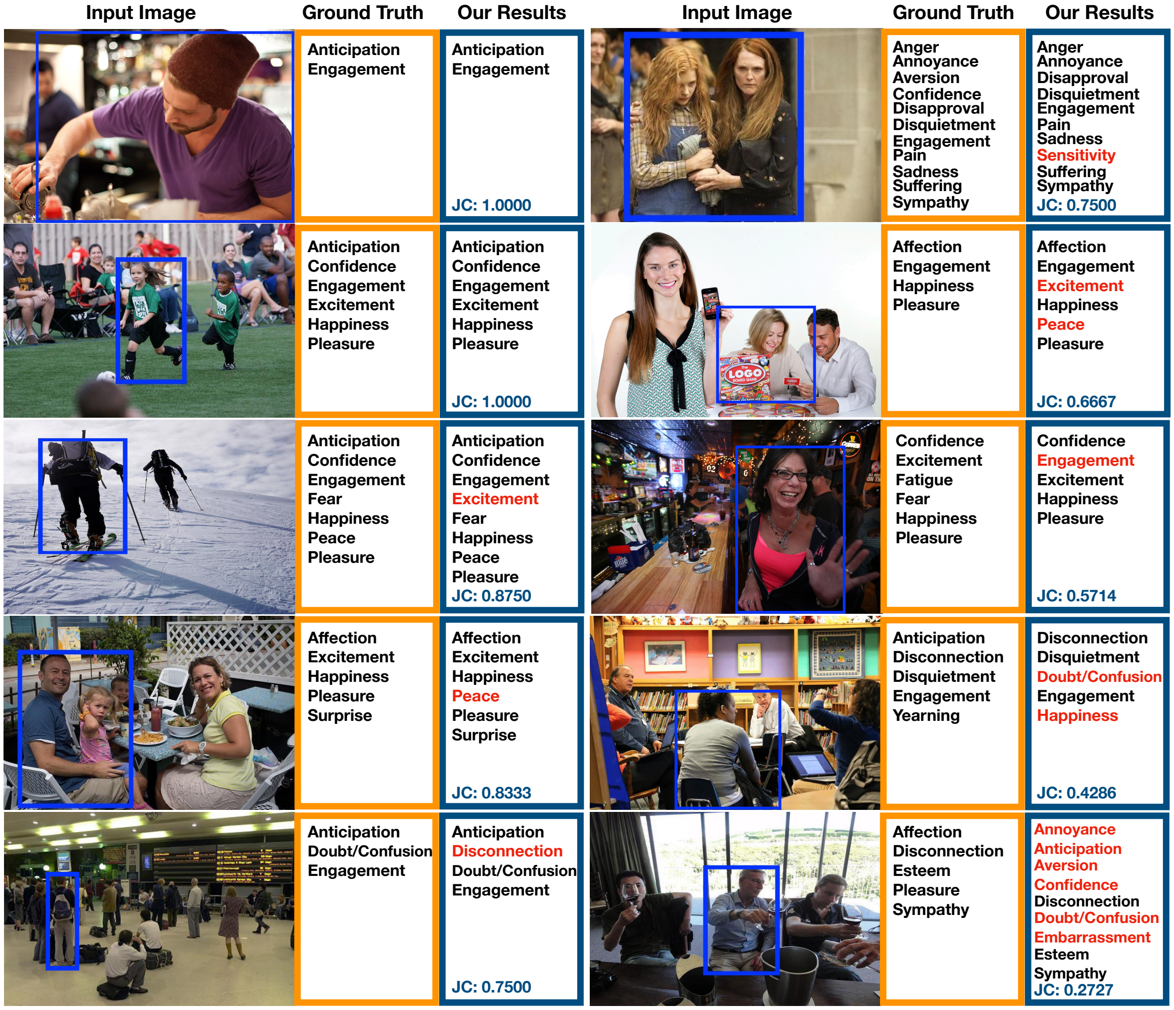


Fig. 8: Qualitative results of our method on the EMOTIC dataset with different Jaccard coefficient (JC) scores (incorrectly inferred emotions are marked in red).

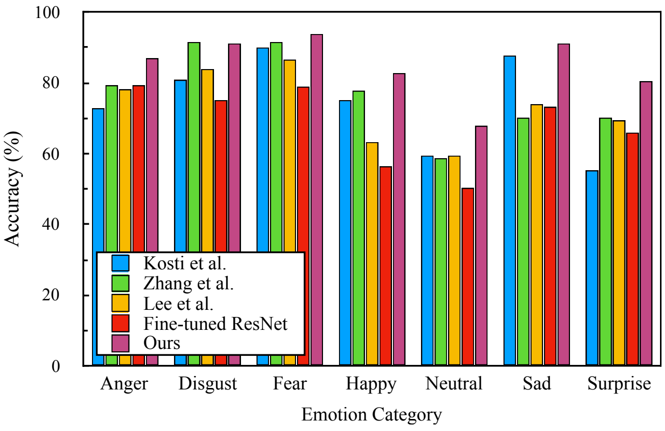


Fig. 9: Emotion classification accuracy for each category of different methods on the CAER-S dataset.

Firstly, we evaluate the contribution of the BOA module. In the methods without BOA, instead of performing the body-object attention, we assign equal weight values to features of all top objects detected by Faster R-CNN. As shown in the first and third method columns in Table 4, we can see that by adding BOA (the third method column), the mAP is improved by 2.65% (compared to the first column). The reason is that even though Faster R-CNN can extract objects within the images effectively, it cannot model the importance of different objects with respect to emotions. In comparison, our BOA module could help attend on regions that are more important for inferring the emotional states of the foreground person. We also evaluate the contribution of geometry features used in the BOA module. The results are shown in the eighth method column in Table 4, where the mAP drops about 0.55% compared to score of the full model in the last column.

Some example results of the BOA module are shown in Fig. 10, in which the emotion-related regions from background scene are displayed along with their weight values describing how they contribute to the emotion perception of the foreground person.

Secondly, we evaluate the contribution of the BPA module. In the methods without BPA, instead of performing the body-part attention module, we directly use the feature of the foreground region extracted by ResNet. By comparing the second method column to the first one in Table 4, we can see that the AP for most emotion categories are improved, and the mAP is also increased by about one percent. The reason mainly lies in that the body-part attention could help refine the features by focusing on parts of the foreground body that are more related to emotions of the person. We also show some example results of the BPA module in Fig. 11. In each row of the figure, the foreground body region

TABLE 4: Ablation study results on the EMOTIC dataset (w/ and w/o are short for with and without, respectively).

	Methods									
BOA	✗	✗	✓	✗	✓	✓	✗	✓(no geo. feats)	✓	✓
BPA	✗	✓	✗	✗	✓	✗	✓	✓	✓(MobileNetV2)	✓
ELD	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓
(1) *Affection*	31.73	32.05	36.66	33.10	37.41	32.62	30.09	38.66	36.81	37.93
(2) *Anger*	9.60	10.45	11.76	11.74	13.71	12.25	11.94	13.38	11.91	13.73
(3) *Annoyance*	11.95	13.66	16.97	17.39	18.22	19.37	17.81	20.11	17.77	20.87
(4) *Anticipation*	59.42	59.72	59.88	59.82	59.17	60.72	59.74	60.26	59.51	61.08
(5) *Aversion*	5.54	6.03	7.80	6.84	8.08	8.30	6.62	8.20	6.03	9.61
(6) *Confidence*	76.70	74.86	79.27	79.78	77.91	79.96	78.59	80.18	73.79	80.08
(7) *Disapproval*	11.99	14.87	15.91	18.80	16.82	20.58	17.92	21.92	17.46	21.54
(8) *Disconnection*	21.99	21.60	25.36	25.15	26.08	26.36	25.70	26.64	25.31	28.32
(9) *Disquietment*	16.13	17.30	19.42	18.77	20.71	21.97	21.82	21.84	19.26	22.57
(10) *Doubt/Confusion*	30.09	31.41	31.22	31.85	33.19	33.32	33.40	32.74	31.00	33.50
(11) *Embarrassment*	2.75	2.85	3.41	2.42	2.77	3.86	4.05	3.53	2.60	4.16
(12) *Engagement*	86.00	85.61	86.85	87.59	86.83	87.89	86.26	87.50	85.34	88.12
(13) *Esteem*	18.73	18.91	18.94	20.33	19.47	19.89	19.60	19.79	18.41	20.50
(14) *Excitement*	77.67	77.38	79.59	80.06	78.23	80.25	79.11	80.14	76.80	80.11
(15) *Fatigue*	12.28	14.14	16.57	13.84	18.19	15.71	15.70	18.08	17.38	17.51
(16) *Fear*	12.11	12.33	13.42	12.46	13.61	14.76	14.31	13.83	11.92	15.56
(17) *Happiness*	65.93	70.81	68.72	74.20	73.76	75.54	74.49	77.29	72.81	76.01
(18) *Pain*	7.33	8.66	10.80	9.42	10.06	14.64	13.86	14.36	9.46	14.56
(19) *Peace*	23.78	24.59	26.81	25.35	26.50	24.37	23.80	27.38	24.36	26.76
(20) *Pleasure*	47.52	48.44	49.88	53.32	52.30	52.85	51.24	54.55	51.99	55.64
(21) *Sadness*	17.43	20.15	23.79	20.73	25.92	26.60	25.01	29.71	23.25	30.80
(22) *Sensitivity*	5.73	6.23	7.48	8.34	6.65	9.38	8.00	8.04	7.08	9.59
(23) *Suffering*	15.16	18.62	22.71	17.85	23.01	28.33	26.70	28.78	21.87	30.70
(24) *Surprise*	17.43	17.36	16.91	17.17	17.18	17.84	18.29	16.96	16.92	17.92
(25) *Sympathy*	13.21	13.55	14.61	13.45	13.89	14.48	14.36	14.84	14.38	15.26
(26) *Yearning*	8.05	8.39	9.14	8.79	9.33	9.41	9.22	9.51	9.07	10.11
mAP	27.11	28.08	29.76	29.56	30.35	31.20	30.29	31.86	29.33	32.41

of the target person is shown along with the heat maps of three top ranked channel features. The results demonstrate that our BPA module can help the network focus more on parts that are helpful for emotion recognition.

As shown in the fifth method column of Table 4, by combining the BOA and BPA modules, the performance of our method is further improved. This illustrates that our relational region-level analysis can effectively account for different contributions of different regions/parts to emotion recognition jointly.

Thirdly, the evaluation of the ELD module is also conducted. We show the performance of our method without and with ELD in the fifth and last method columns in Table 4. As shown, the AP scores of almost all emotion categories are improved by modeling the dependency between different emotions. And some emotion categories benefit a lot from the ELD module, especially some “negative” emotion categories, e.g. *Pain*, *Sadness*, and *Suffering*. We also show some example results of the ELD module in Fig. 12. Emotions inferred without and with the module are shown in the figure, and the results also demonstrate qualitatively that modeling emotion label dependency can contribute to more accurate emotion recognition.

Lastly, we change the baseline model used in the proposed method before the BPA module, i.e. ResNet50, to a popular lightweight model MobileNetV2 [63]. The results are shown in the ninth method column in Table 4. As we can see from the table, even though the lightweight models are more efficient, their recognition accuracy still gets sacrificed. We will explore how to design an effective and efficient emotion recognition model in the future work, if we need to make our method suitable for the potential applications on computation-limited platforms.

We also evaluate the choices of some parameter values used in the experiment. The head number h is a parameter for the multi-head attention in the BOA module. The mAP scores on the EMOTIC datasets for models with $h = 4, 16, 32$ are 31.78%, 32.4%, and 32.16%, respectively. So the change of head number does not have a huge impact on the model performance, and we choose to set h=16 in the experiments for other model variants. p is the parameter for determining the weights assigned to a node with respect to other correlated nodes in the ELD module, as shown in Eq. 7. Changing the value of p in the set $\{0, 0.1, 0.2, \ldots, 0.9, 1\}$ results in the change of the mAP scores in the range of 29.53% to 32.41%. And we choose to set $p = 0.2$ where the highest mAP score is achieved.

Moreover, we evaluate the proposed method with/without the ELD module on the Acted Facial Expressions in the Wild (AFEW) dataset [64]. AFEW dataset contains movie video clips that capture facial expressions annotated with seven basic emotions. The whole dataset is split into training (773 video clips), validation (383), and test (653) sets. In the experiment, we follow [65] to evaluate our method using the validation set since ground-truth labels for the test set are not available, and the final emotion classification result for each clip is calculated based on the summation of predicted emotion scores of all frames. To evaluate our method without ELD module, we set the adjacency matrix $\mathbf{A}$

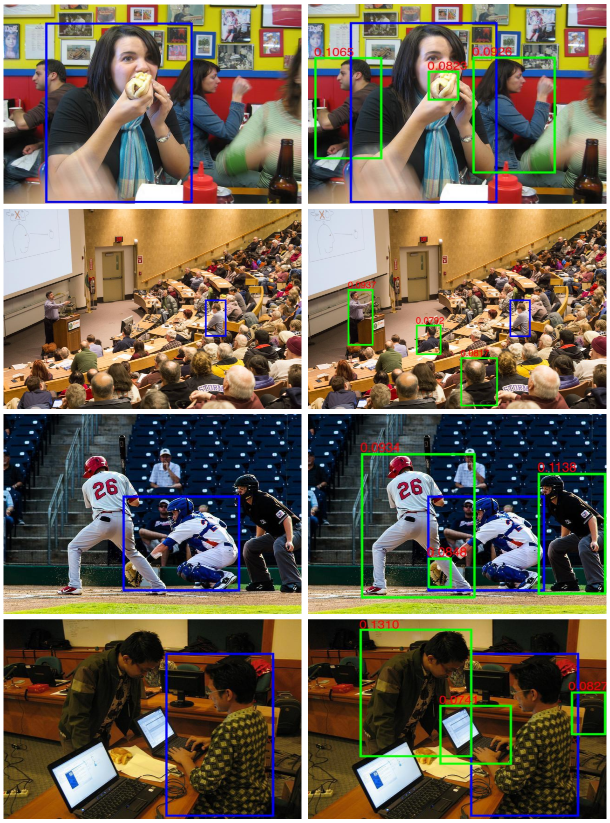


Fig. 10: Example results of the BOA module. The left column shows the input image and the target person. The right column shows the top-3 attended regions (marked in green) and their weight values (marked in red). These examples show that, given the target person, the BOA module can help attend on emotion-related regions.

as an identity matrix.

The emotion recognition accuracies of our method with and without the ELD module on AFEW dataset are 42.04% and 42.30%, respectively. Compared to the ResNet50 model [65] which has the recognition accuracy of 40.99%, our method has better performance. Since in our general emotion recognition framework, no specific module is designed for accurately modeling facial expressions, our performance is not that satisfactory compared to the ResNet50 with deeply-supervised blocks (DSN-ResNet50) proposed in [65], which has the recognition accuracy of 43.86%. And since video clips in the AFEW dataset are only labeled with 7 basic emotions, the ELD module is not beneficial as expected. And the performance of our method slightly drops when adding the ELD module, probably due to the increased model complexity.

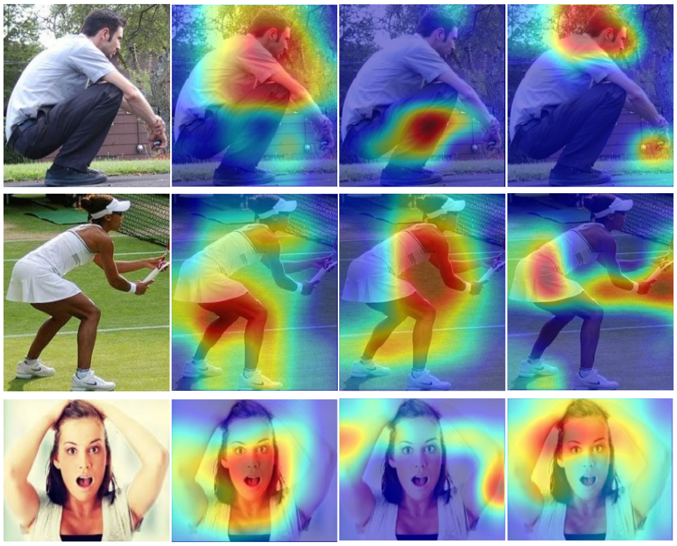
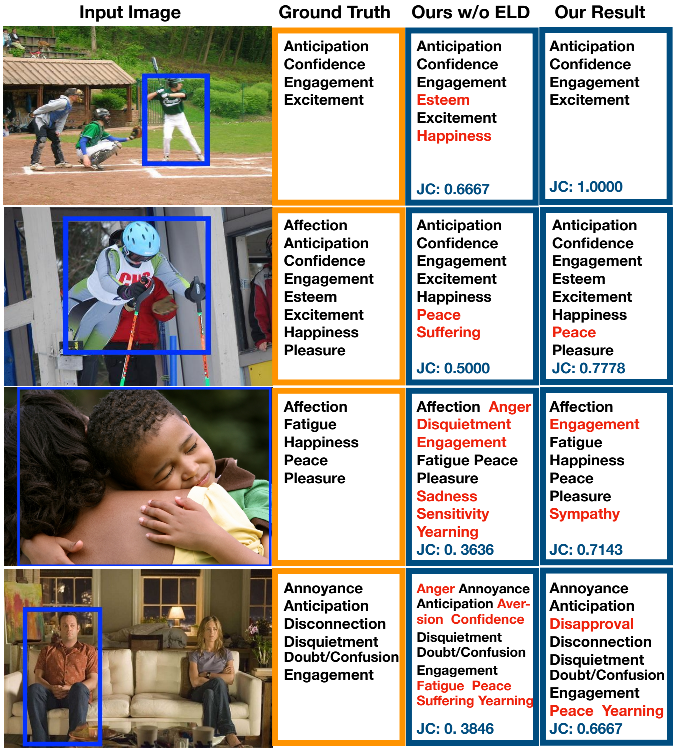


Fig. 11: Example results of the BPA module. From left to right are the foreground body region of the target person, and heat maps of three top ranked channel features. These examples show that the BPA module can help the model focus on body parts that are related to emotion perception of the target person.



Input Image	Ground Truth	Ours w/o ELD	Our Result
	Anticipation Confidence Engagement Excitement	Anticipation Confidence Engagement Esteem Excitement Happiness JC: 0.6667	Anticipation Confidence Engagement Excitement JC: 1.0000
	Affection Anticipation Confidence Engagement Esteem Excitement Happiness Pleasure	Anticipation Confidence Engagement Excitement Happiness Peace Suffering JC: 0.5000	Anticipation Confidence Engagement Esteem Excitement Happiness Peace Pleasure JC: 0.7778
	Affection Fatigue Happiness Peace Pleasure	Affection Anger Disquietment Engagement Fatigue Peace Pleasure Sadness Sensitivity Yearning JC: 0. 3636	Affection Engagement Fatigue Happiness Peace Pleasure Sympathy JC: 0.7143
	Annoyance Anticipation Disconnection Disquietment Doubt/Confusion Engagement	Anger Annoyance Anticipation Aversion Confidence Disquietment Doubt/Confusion Engagement Fatigue Peace Suffering Yearning JC: 0. 3846	Annoyance Anticipation Disapproval Disconnection Disquietment Doubt/Confusion Engagement Peace Yearning JC: 0.6667

Fig. 12: Example results of the inferred emotions without and with the ELD module (incorrect emotions are marked in red, and w/o is short for without). These examples demonstrate that by modeling the label dependency, more accurate emotion recognition can be achieved.

5 Conclusion

We have presented an end-to-end convolutional neural network for emotion recognition in real-world scenarios based on relational region-level analysis. 1) For objects in the background scene, we propose the body-object attention module to estimate contributions of the background objects to the emotion recognition of the foreground body, based on the appearance and geometry features of these regions. 2) For the foreground body, we propose the body part attention module, which refines the channel-wise body features to focus on emotion-related body parts. 3) The dependency across different emotion labels is also exploited and modeled based on both their semantic meanings and co-occurrence patterns with Graph Convolutional Network. Experiments on two datasets for emotion recognition in real-world scenarios, i.e. the EMOTIC dataset and CAER-S dataset, show that our method can achieve superior performance compared to the state-of-the-art algorithms.

Acknowledgment

This work is funded by the National Nature Science Foundation of China (No. 61802026 and 61806016) and the Fundamental Research Funds for the Central Universities (No. 2019RC39). This work is also sponsored by CAAI-Huawei MindSpore Open Fund.

References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[3] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[4] M. Alhussein, "Automatic facial emotion recognition using weber local descriptor for e-healthcare system," *Cluster Computing*, vol. 19, no. 1, pp. 99–108, 2016.

[5] F. Wang, M. Mao, L. Duan, Y. Huang, Z. Li, and C. Zhu, "Intersession instability in fNIRS-based emotion recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1324–1333, 2018.

[6] M. Čertickỳ, M. Čertickỳ, P. Sinčák, G. Magyar, J. Vaščák, and F. Cavallo, "Psychophysiological indicators for modeling user experience in interactive digital entertainment," *Sensors*, vol. 19, no. 5, p. 989, 2019.

[7] P. A. Nogueira, R. Rodrigues, E. Oliveira, and L. E. Nacke, "Guided emotional state regulation: Understanding and shaping players' affective experiences in digital games," in *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.

[8] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2168–2177.

[9] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.

[10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.

[11] D. H. Nguyen, S. Kim, G. Lee, H. Yang, I. Na, and S. H. Kim, "Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks," *IEEE Transactions on Affective Computing*, 2019.

[12] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, 2019.

[13] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 369–381, 2017.

[14] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[15] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[16] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[17] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 10 143–10 152.

[18] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017, pp. 1–10.

[19] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1960–1968.

[20] Z. Li, J. Imai, and M. Kaneko, "Facial-component-based bag of words and PHOG descriptor for facial expression recognition," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 1353–1358, 2009.

[21] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[22] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2584–2593.

[23] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5562–5570.

[24] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2562–2569.

[25] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[26] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[27] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," in *International Conference on Artificial Intelligence for Human Computing*, 2007, pp. 91–112.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[29] K. Schindler, L. V. Gool, and B. Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008.

[30] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, no. C, pp. 140–151, 2015.

[31] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, 2017.

[32] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Computer Vision and Image Understanding*, vol. 174, pp. 33–42, 2018.

[33] Z. Chen and D. Whitney, "Tracking the affective state of unseen persons," *Proceedings of the National Academy of Sciences*, vol. 116, no. 15, pp. 7559–7564, 2019.

[34] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 151–156.

[35] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-aware multimodal emotion recognition using Frege's principle," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 234–14 243.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[38] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3588–3597.

[39] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[40] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2285–2294.

[41] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of*

the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 464–472.

[42] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2027–2036.

[43] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2977–2986.

[44] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5172–5181.

[45] S. Ruan, K. Zhang, Y. Wang, H. Tao, W. He, G. Lv, and E. Chen, "Context-awar generation-based net for multi-label visual emotion recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[46] J. Yang, D. She, M. Sun, M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[47] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017, pp. 231–237.

[48] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.

[49] A. Gonzalez-Garcia, D. Modolo, and V. Ferrari, "Do semantic parts emerge in convolutional neural networks?" *International Journal of Computer Vision*, vol. 126, no. 5, pp. 476–494, 2018.

[50] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6995–7003.

[51] M. Simon, E. Rodner, and J. Denzler, "Part detector discovery in deep convolutional neural networks," in *Asian Conference on Computer Vision (ACCV)*, 2014, pp. 162–177.

[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[53] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[54] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-Label Image Recognition with Graph Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5177–5186.

[55] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 10 143–10 152.

[56] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genetic, Social, and General Psychology Monographs*, vol. 121, no. 3, pp. 339–361, 1995.

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

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[60] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "MAttNet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1307–1315.

[61] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

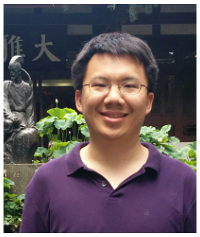
[62] R. Kosti, "Visual scene context in emotion perception," Ph.D. dissertation, Universitat Oberta de Catalunya, 2019.

[63] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018.

[64] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[65] Y. Fan, J. C. Lam, and V. O. Li, "Video-based emotion recognition using deeply-supervised neural networks," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 584–588.

Weixin Li received the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2017. She is currently an Associate Researcher at the School of Computer Science and Engineering (SCSE) and Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, Beijing, China. Her research interests include computer vision, image processing, and big data analytics.



Xuan Dong received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2015, and the B.E. degree in Computer Science from Beihang University, Beijing, China, in 2010. He is currently an Associate Professor at the School of Computer Science, Beijing University of Posts and Telecommunications, China. His research interests include computer vision and computational photography.

Yunhong Wang received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively. She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her current research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.