Regional Attention Networks with Context-aware Fusion for Group Emotion Recognition

Ahmed Shehab Khan[1], Zhiyuan Li[1], Jie Cai[2], Yan Tong[1]
[1]Department of Computer Science & Engineering, University of South Carolina, Columbia, SC
[2]InnoPeak Technology, Palo Alto, CA
{akhan,zhiyuanl}@email.sc.edu, jie.cai@innopeaktech.com, tongy@cec.sc.edu

Abstract

Group Emotion Recognition (GER) from images has many inherent challenges. Specifically, it is difficult to combine diverse emotions of different individuals into a single conclusive label. In addition, although utilization of information other than faces like scene and objects has proven helpful, it is still a challenge to effectively fuse predictions of individual sources. In this work, we proposed solutions to these two problems. First, we developed a regional attention mechanism to find important persons or objects, which play critical roles in the group emotion, and combine them based on importance. Second, we proposed a context-aware fusion mechanism to estimate weights from the image context to fuse different sources of information. Finally, we proposed to use a single backbone network to extract features from multiple sources, i.e., scene, faces, and objects, cutting down computation and memory cost. Experiments on two GER datasets have shown that the proposed framework achieves performance comparable to the state-of-the-art. Furthermore, a visualization study and a case study have demonstrated that the proposed model is effective to extract and more importantly, emphasize the most critical information in GER.

1. INTRODUCTION

Emotions play a crucial role in our everyday life and impact the way we communicate and interact with others. Being empathetic, we are highly responsive to the surrounding environment; and as we constitute the environment, we largely influence it by our interaction. A vast majority of the research conducted in emotion recognition (ER) is done on Individual Emotion Recognition (IER), i.e, recognizing the emotion of a single individual. Recently, thanks to the availability of images on the web and social media, the frontier of emotion recognition research has been pushed by a big margin, and more importantly, an opportunity emerges to advance knowledge in Group Emotion Recognition (GER), which aims to understand the emotion from a group of people. In addition to dealing with existing challenges of IER, such as head pose variation, occlusion, and racial difference among people, GER has its own complexity.

Figure 1. a) and b) are examples of diverse expressions in the same image. Although most of faces show a neutral expression in c) and d), the presence of objects like banners and signboards suggests a negative emotion in c), whereas the clues from objects like the cake, balloons, and hats imply a positive event.

First, it is not straightforward to understand the group emotion from multiple individuals, since each person in an image may not show the same expression. For example, in Fig. 1 a) and b), we can observe that the individuals involved are displaying different and in some cases, even opposite expressions. In Fig. 1 a), the sad face of the kid makes the dominant impact on the group emotion, even though one of the ladies smiles. Similarly, in Fig. 1 b), the smiling face of the lady overshadows the angry-looking expression of the gentleman. One more observation in this image is that the persons in the background do not belong to the group and thus, should be ignored in GER. Similar is the case for objects involved. For example, banners and posters in Fig. 1 c) give a negative vibe for the group, whereas the cars in the

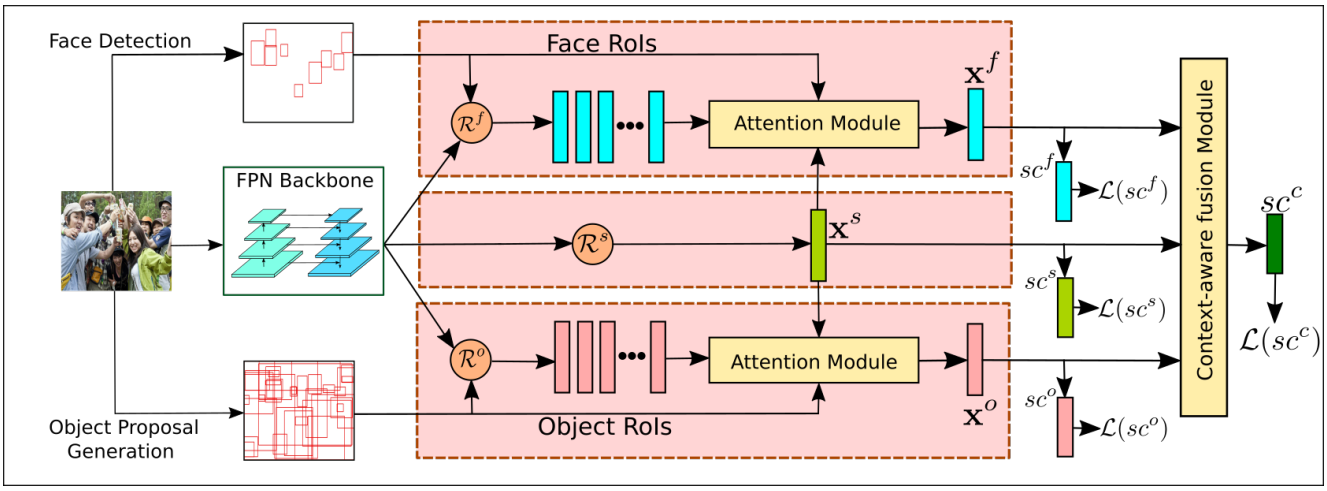


Figure 2. An overview of the proposed GER framework, where images are passed through a shared backbone FPN network and then are fed to three streams. $\mathcal{R}^s$, $\mathcal{R}^f$, and $\mathcal{R}^o$ are RoI feature extractors for the scene, face, and object streams, respectively. $\mathcal{L}(sc^s)$, $\mathcal{L}(sc^f)$, $\mathcal{L}(sc^o)$, and $\mathcal{L}(sc^c)$ are loss terms associated with the three streams and the final prediction via fusion, respectively.

background are irrelevant, making it challenging to determine objects with the most information. To mitigate this challenge, we proposed a regional attention mechanism to differentiate and more importantly, estimate the importance of persons or objects in group emotion.

Second, while gathering information from multiple streams like scene, faces, and objects has shown to be effective in improving GER performance [13, 32, 18], it remains challenging to combine the predictions of all streams into a single decision, since the importance of different streams may vary according to the context. For example, while we rely more on the face stream in Fig. 1 a) and b), much less information can be extracted from the faces in c) and d). Instead, objects like banner and posters in Fig. 1 c) or the cake and hats in Fig. 1 d) provide strong clues in determining group emotion. To address this challenge, we proposed a context-aware fusion mechanism to estimate fusion weights of multiple streams from the image content.

Third, recent approaches of GER [12, 13, 32, 18] takes the advantage of multiple streams of information, while each stream employs a separate network. Since all the streams are derived from the same image, it is overkill and computationally expensive to use separate networks for each stream. To this front, we proposed, for the first time, to use a single shared backbone network for all streams.

Fig. 2 depicts an overview of the proposed GER framework. Specifically, images are passed through a backbone network based on Feature Pyramid Network (FPN) [21], from which feature maps are generated and shared by three streams. A scene stream extracts scene features $\mathbf{x}^s$ from the whole feature maps. Given Regions of Interests (RoIs), i.e., face bounding boxes, FPN feature maps, and the scene features $\mathbf{x}^s$, a face stream extracts face features $\mathbf{x}^f$ through the proposed regional attention module. The object features $\mathbf{x}^o$ are extracted similarly through the object stream. $\mathbf{x}^s$, $\mathbf{x}^f$, and $\mathbf{x}^o$ are then combined by the proposed context-aware fusion module to obtain the final decision score sc^c.

In summary, the major contributions of the paper are as follows:

- A regional attention mechanism, to estimate the importance of a person or an object in the context of the scene and to aggregate information accordingly,
- A context-aware fusion mechanism that learns fusion weights of multiple streams from the image, and
- A GER framework that employs a single shared backbone network as the feature extractor to handle scale variations and to save computation and memory cost.

2. RELATED WORK

Individual Emotion Recognition (IER), a precursor to the GER task, has been vastly studied over the last decade and significantly improved the ability that computers can understand the emotion of an individual. Lately, deep learning based approaches have seen significant progress [3, 25, 22, 2, 9, 10, 15] in ER. Although most of these approaches explicitly analyze the facial region to determine emotion, several attempts were made to use signals from other sources to enhance the accuracy of ER [29, 26, 4, 19, 34, 20]. For example, Nicolaou et al. [26] used location of shoulders and Schindler et al. [29] used body pose to enhance emotion recognition. Various other approaches utilized context information [4, 19, 34, 20]. For instance, Chen et al. [4] used pre-trained CNNs to generate scores from events, objects, and scene, and then used these scores as features to train a neural network. Kosti et al. [19] extracted features from body pose and scene, and combined them to predict emotion of an individual. Lee et al. [20] used a face encoding stream and a context encoding stream

and fused them with learned weights.

The success of these approaches provides a strong indication of the importance of context on IER and inspired our proposed method. However, these approaches are mostly focused on a single person or the scene, and do not explicitly consider the presence of multiple people in a group. Moreover, they predict the emotion of an individual, but not for the group as a whole, which requires an understanding of how much an individual is contributing towards the group.

Group Emotion Recognition (GER) is the task of determining emotion from a group of individuals. Unfortunately, this problem has not been well studied in the past, mostly due to unavailability of data. Dhall et al. [7] presented a Multiple Kernel Learning based hybrid GER inference model and published the Group Affect Database, which contains images of a group of people at a social event labeled as "Positive", "Negative" or "Neutral". Later, EmotiW group-level emotion recognition sub-challenge was initiated to advance GER task [6, 8]. As a positive outcome of the challenge, several attempts have been made to solve this task [30, 11, 13, 32, 18, 14].

In addition to faces and scene, some approaches employed more sources of information. Khan et al. [18] used face location information in form of attention heatmaps. Wang et al. [32] used human body. Guo et al. [13] utilized skeletons and objects. Guo et al. further developed a graph neural network based approach [12] considering the interactions between various nodes, where the nodes are features extracted from several streams, i.e, faces, objects, human patches, and the scene.

All these aforementioned approaches trained a separate network for individual streams, predictions of which are later fused for final classification. In contrast, we proposed to use a single shared backbone FPN, considering a number of advantages. First, since the first few layers learn low-level features [36], which are similar for all networks, sharing backbone can cut down memory usage and computational cost without sacrificing performance. Second, as the backbone is input size agnostic, RoIs do not need to be reduced to a fixed smaller size, and thus not lose critical information. Third, RoIs like faces in the GER task have a large scale variation [18], and hence, usage of FPN gives the ability to explicitly consider different scales. Finally, training of the whole network can be performed end to end optimizing a single loss function, as opposed to training multiple networks separately with different loss functions.

In order to integrate different streams, several fusion schemes have been adopted by the aforementioned approaches. For example, Gupta et al. [14] used concatenated features from face and scene streams. Weighted-averaging has been used in [18, 13, 32], where the weights are learned from exhaustive grid search on the validation set. Guo et al. [12] used majority voting for final prediction. In all of these approaches, the weights are fixed after the training and do not change with the context of the image. Considering large variations in image context of GER as discussed earlier, we proposed a context-aware fusion mechanism to learn weights explicitly from image content.

Visual attention has been widely used and shown enormous success in many areas including image captioning [1, 23, 28], visual question answering [33, 35], image classification [31], and image generation [37]. Attention can be used to find the relative importance of a set of contextual regions and has also been applied in the GER task [14, 32]. Different from these methods, which learn attention only from appearance features, we proposed to employ geometric information of RoIs as an additional signal as well as global scene features extracted from the same backbone to calculate the attention of the contextual region.

3. METHODOLOGY

An overview of the proposed GER framework is illustrated in Fig. 2. First, face bounding boxes and object proposal bounding boxes are detected using an off-the-shelf face detector [38] and an object proposal network [1], respectively. An input image is then passed through a shared backbone network, which produces intermediate feature maps. From the shared backbone network stems three independent streams. The scene stream extracts features from the whole scene denoted by $\mathbf{x}^s$. The face stream extracts face features denoted by $\mathbf{x}^f$ using the proposed regional attention module. Similarly, the object stream extracts object features denoted by $\mathbf{x}^o$. Finally, the features from all three streams, i.e., $\mathbf{x}^s$, $\mathbf{x}^f$, and $\mathbf{x}^o$, are combined through the proposed context-aware fusion module to determine the final classification score. Details of each component are described in the following.

3.1. RoI Feature Extraction from A Shared Backbone Network

In this work, faces are detected by MTCNN [38], which is a deep cascaded multi-task framework for face and landmark detection. The detected face bounding boxes are used as the RoIs for the face stream and the ith detected face is denoted by b_i^f. For the object stream, the RoIs are object proposals. Following [12], we use a pre-trained bottom-up-attention network [1] for object proposal generation and denote the ith object RoI by b_i^o. For the scene stream, we use a single bounding box with the same size of the image as the RoI and denote it by b^s.

Motivated by the fact that all the streams derive from the same image, we proposed to use a single shared backbone network to extract intermediate features for all streams. Specifically, a 50-layer Deep Residual Network (Resnet) [17] with Feature Pyramids [21], i.e., Resnet-50-FPN, is used as the backbone network, producing feature maps denoted by $\mathbf{B}$.

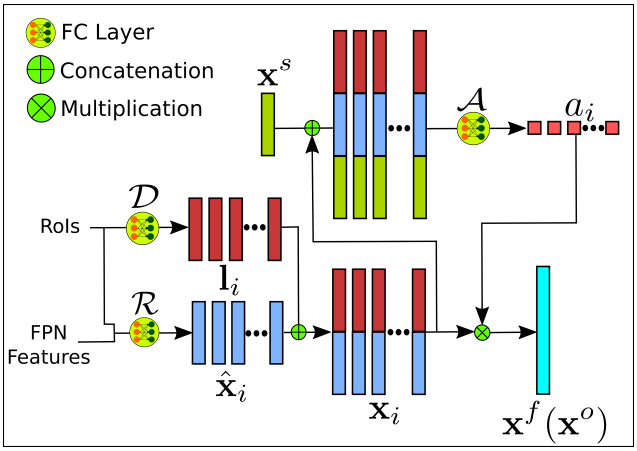


Figure 3. Illustration of the feature extraction process for the face/object stream given the input of FPN feature maps and RoIs.

To facilitate feature-level fusion from multiple RoIs, RoIAlign [16] is employed to extract fixed-sized features from the feature maps to handle scale variations in RoIs. In addition, since we have feature maps of different scales from the FPN, the correct feature map to extract features is determined by the same heuristic as [21]. Then, appearance features corresponding to each RoI are extracted as follows:

$$\begin{aligned} \mathbf{x}^s &= \mathcal{R}^s(RoIAlign(b^s, \mathbf{B})) \\ \hat{\mathbf{x}}_i^f &= \mathcal{R}^f(RoIAlign(b_i^f, \mathbf{B})) \\ \hat{\mathbf{x}}_i^o &= \mathcal{R}^o(RoIAlign(b_i^o, \mathbf{B})) \end{aligned} \tag{1}$$

where $RoIAlign(\cdot)$ takes in a bounding box of an RoI and feature maps from the backbone $\mathbf{B}$ and returns a fixed-size feature vector. As shown in Fig. 3, $\mathcal{R}^s(\cdot)$, $\mathcal{R}^f(\cdot)$, and $\mathcal{R}^o(\cdot)$ are feature extractor functions for the scene, face, and object streams, respectively. Each feature extractor is implemented as three fully connected (FC) layers with 1024, 1024, and 128 neurons, respectively. It should be noted that these feature extractors are independent and do not share weights. $\mathbf{x}^s$, $\hat{\mathbf{x}}_i^f$, and $\hat{\mathbf{x}}_i^o$ represent the extracted scene feature vector, the appearance feature vector of the ith face RoI, and the appearance feature vector of the ith object proposal, respectively.

However, appearance features extracted by the RoI feature extractor completely lose spatial information of the RoI and the relative positions to other RoIs. To address this issue, geometric information is explicitly added to the RoI features. Since the geometric features consist of only eight elements, which are insignificant compared to the 128-D appearance features, they are further up-sampled to have the same size as the appearance features as below:

$$\begin{aligned} \mathbf{l}_i^f &= \mathcal{D}(c_{x_i}^f, c_{y_i}^f, w_i^f, h_i^f, S_i^f, \hat{w}_i^f, \hat{h}_i^f, \hat{S}_i^f) \\ \mathbf{l}_i^o &= \mathcal{D}(c_{x_i}^o, c_{y_i}^o, w_i^o, h_i^o, S_i^o, \hat{w}_i^o, \hat{h}_i^o, \hat{S}_i^o) \end{aligned} \tag{2}$$

where c_{x_i}, c_{y_i}, w_i, h_i, and S_i represent the x and y coordinates of the center, width, height, and area of the ith RoI, normalized by the image size; $\hat{w}_i$, $\hat{h}_i$, and $\hat{S}_i$ represent the width, height, and area of the ith RoI, normalized by the maximum width, height, and area among all the RoIs. The superscripts f and o represent the face and object streams, respectively. $\mathcal{D}(\cdot)$ is an up-sampling function implemented by an FC layer with 128 output neurons and shares weights for both face and object streams.

As shown in Fig. 3, each RoI is represented by concatenating appearance features from Eq. (1) and geometric features from Eq. (2) as follows:

$$\mathbf{x}_i^f = \hat{\mathbf{x}}_i^f \oplus \mathbf{l}_i^f \ , \ \mathbf{x}_i^o = \hat{\mathbf{x}}_i^o \oplus \mathbf{l}_i^o \tag{3}$$

where $\mathbf{x}_i^f$ and $\mathbf{x}_i^o$ are complete feature vectors of the ith face RoI and the ith object RoI, respectively. $\oplus$ is concatenation operation.

3.2. Regional Attention Module

Since there are multiple RoIs in the face/object stream, we propose a regional attention module to perform feature-level fusion for the face/object stream. As shown in Fig. 3, the attention module takes the input of both appearance and geometric information of each RoI as well as the global context, i.e., the scene feature $\mathbf{x}^s$, to determine the importance of each RoI as below:

$$\begin{aligned} \hat{a}_i^f &= \mathcal{A}^f(\mathbf{x}^s \oplus \mathbf{x}_i^f) \ , \ a_i^f = \frac{exp(\hat{a}_i^f)}{\sum_j^{N_f} exp(\hat{a}_j^f)} \\ \hat{a}_i^o &= \mathcal{A}^o(\mathbf{x}^s \oplus \mathbf{x}_i^o) \ , \ a_i^o = \frac{exp(\hat{a}_i^o)}{\sum_j^{N_o} exp(\hat{a}_j^o)} \end{aligned} \tag{4}$$

where a_i^f and a_i^o represent the importance of the ith face RoI and the ith object RoI, respectively. $\mathcal{A}(\cdot)$ is the attention function, which is implemented by an FC layer with 1 output neuron.

Once the importance of each RoI is estimated, features of multiple RoIs are combined together to form a single feature vector for each stream:

$$\mathbf{x}^f = \frac{1}{N_f}\sum_{i=1}^{N_f} a_i^f * \mathbf{x}_i^f \ , \ \mathbf{x}^o = \frac{1}{N_o}\sum_{i=1}^{N_o} a_i^o * \mathbf{x}_i^o \tag{5}$$

where $\mathbf{x}^f$ and $\mathbf{x}^o$ are the final feature representations of the face and object streams, respectively; N_f and N_o are the numbers of face RoIs and object RoIs, respectively.

3.3. Context-aware Fusion Module

Given the extracted feature vector for each stream, we can build a classifier to get a score for GER as follows:

$$sc^s = \mathcal{F}^s(\mathbf{x}^s) \ , \ sc^f = \mathcal{F}^f(\mathbf{x}^f) \ , \ sc^o = \mathcal{F}^o(\mathbf{x}^o) \tag{6}$$

where sc^s, sc^f, and sc^o are classification scores for the scene, face, and object stream, respectively. $\mathcal{F}^s(\cdot)$, $\mathcal{F}^f(\cdot)$,

and $\mathcal{F}^o(\cdot)$ are classifiers, each of which is implemented by a FC layer with 3 neurons, representing three emotional categories, i.e., positive, negative, and neutral.

The widely adopted score-level fusion strategy is to combine these scores by a weighted average, where weights can be learned empirically from the validation set and then fixed during testing. However, as shown in Fig. 1, the contributions of different streams to GER highly depend on the image context. For example, Fig. 1 a) and b) contain less background information, but have clear faces with strong expressions. Hence, the face stream should be dominant in determining the group emotion. On the contrary, objects in Fig. 1 c) and d) give strong emotional cues, whereas the performance of face stream is significantly impaired by small face size, occlusion, and large face pose, which are inherent challenges in facial expression recognition.

Motivated by this observation, we propose to learn the fusion weights from the image itself by a context-aware fusion module. Specifically, the inputs to the context-aware weight function are the feature vectors extracted from the scene, face, and object streams respectively.

$$\hat{\mathbf{W}} = \mathcal{G}(\mathbf{x}^s \oplus \mathbf{x}^f \oplus \mathbf{x}^o) = [\hat{\mathbf{w}}^s, \hat{\mathbf{w}}^f, \hat{\mathbf{w}}^o]$$
$$W_{i,j} = \frac{\exp(\hat{W}_{i,j})}{\sum^{k=3} \exp(\hat{W}_{i,k})} \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{3\times 3}$, of which rows represent the emotion categories, i.e, positive, negative, and neutral, and columns represent the streams, i.e, scene, face, and object. $\mathcal{G}(\cdot)$ is the context-aware weight function implemented by an FC layer with 9 output neurons. $\mathbf{W}$ is normalized such that the weights of three streams for each emotion category sum to 1. $[,]$ is a channel-wise concatenation operator.

Finally, scores of the different streams are combined to a final classification score sc^c based on the weight matrix $\mathbf{W}$ learned from the image itself.

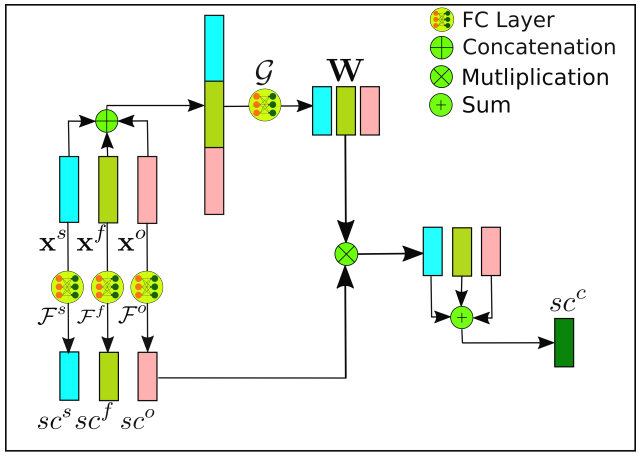


Figure 4. Illustration of the proposed context-aware fusion.

3.4. Governing Loss Function

The proposed GER framework can be trained end-to-end except the standalone face detector and object proposal network. During training, we have four loss terms: $\mathcal{L}(sc^c)$ represents the final prediction and the others represent individual predictions from each stream. The overall loss function to guide the whole training process is as below:

$$\mathcal{L}oss = \lambda_c \mathcal{L}(sc^c) + \lambda_s \mathcal{L}(sc^s) + \lambda_f \mathcal{L}(sc^f) + \lambda_o \mathcal{L}(sc^o) \tag{8}$$

where $\mathcal{L}(\cdot)$ is a sigmoid cross-entropy loss function; λ_c, λ_s, λ_f, and λ_o are the contribution weights towards the overall $Loss$. In this work, these four weights are all set to 1.

4. EXPERIMENTS

4.1. Experimental Datasets

GroupEmoW [12] dataset contains 15,894 images and is divided into "Train", "Validation" and "Test" subsets with $11,127$, $3,178$ and $1,589$ images, respectively. The images of this dataset were collected from the web by searching in popular engines, e.g., Google, Baidu, Bing, and Flickr, with keywords related to social events, such as funeral, birthday, protest, conference, meeting, etc. Each image is labeled to one of "Neutral", "Positive", and "Negative" states. The annotation task was performed by multiple persons, and then the ground truth label was determined by consensus. Ground truth labels for all "Train", "Validation", and "Test" subsets have been made publicly available.

Group Affect Database 2.0 [7] consists of $17,172$ images and is divided into three subsets, i.e., "Train", "Validation", and "Test", containing $9,815$, $4,346$ and $3,011$ images, respectively. The images were collected from Google and Flickr using keywords corresponding to different events, e.g., festival, party, silent protest, violence, etc. Each image in the dataset belongs to one of the three classes: "Neutral", "Positive", and "Negative". This dataset was used for EmotiW2018 GER sub-challenge [8]. The ground truth labels of "Train" and "Validation" subsets are publicly available, while "Test" subset is kept closed by the organizers.

4.2. Implementation Details

Preprocessing: For the face stream, face bounding boxes were detected using MTCNN [38]. For the object stream, a bottom-up-attention method [1] was used to generate proposal bounding boxes, among which the first 36 proposals were selected based on confidence, as an input to the GER network. For data augmentation purpose, we randomly cropped and horizontally flipped the input images with 40% probability.

Training Strategy: The shared backbone Resnet-50-FPN was initialized following maskrcnn-benchmark [24], which

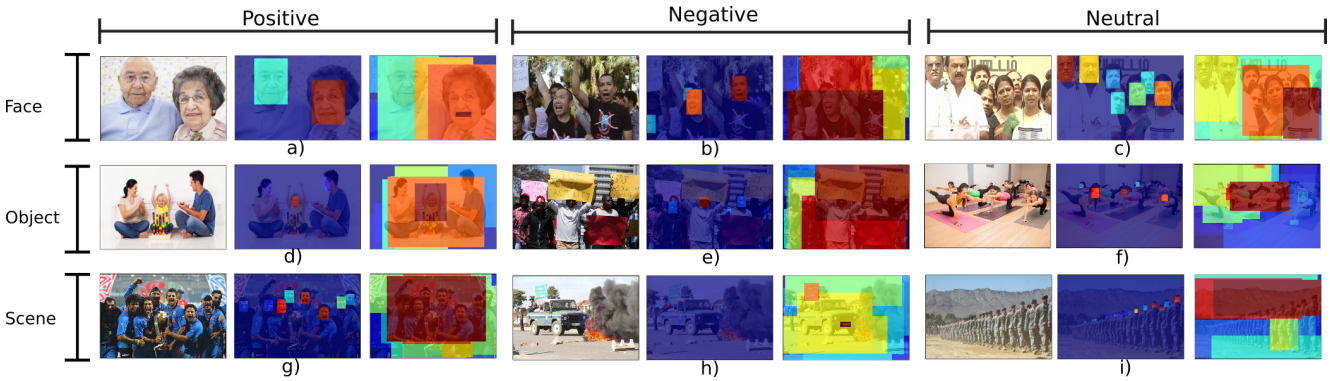


Figure 5. Visualization of attention heatmaps. The rows represent the stream having the highest weight in context-aware fusion with three examples of "Positive", "Negative", and "Neutral". Each example has three depictions for the original image, the face stream attention heatmap, and the object stream attention heatmap, respectively.

Table 1. Experimental results on Group Affect database 2.0 in terms of overall recognition accuracy.

Method	Validation	Test	Sources
Inception-Img [8]	65.0	61.00	Scene
SE-ResNet-50 [13]	68.16	-	Scene
Khan et al. [18]	78.39	66.29	Scene, Faces
Wang et al. [32]	**86.90**	67.49	Scene, Faces, Human body
Guo et al. [13]	78.98	**68.08**	Scene, Faces, Objects, Skeleton
GNN [12]	79.08	-	Scene, Faces, Objects, Skeleton
BL_S	77.88	65.52	Scene
BL_{SF}	78.14	67.18	Scene, Faces
BL_{SFO}	78.46	67.38	Scene, Faces, Objects
RAN	78.76	67.08	Scene, Faces, Objects
CARAN	**79.13**	**67.61**	Scene, Faces, Objects

Table 2. Experimental results on GroupemoW database in terms of overall recognition accuracy.

Method	Test	Sources
SE-ResNet-50 [12]	82.38	Scene
GNN [12]	89.93	Scene, Faces, Objects
BL_S	88.29	Scene
BL_{SF}	89.36	Scene, Faces
BL_{SFO}	89.61	Scene, Faces, Objects
RAN	90.02	Scene, Faces, Objects
CARAN	**90.18**	Scene, Faces, Objects

was trained on image classification task on ImageNet [5]. The remaining network parameters, i.e., parameters of $\mathcal{D}(\cdot)$, $\mathcal{R}(\cdot)$, $\mathcal{A}(\cdot)$, $\mathcal{G}(\cdot)$, and $\mathcal{F}(\cdot)$, were initialized from a uniform distribution with a mean 0 and a standard deviation of 0.01. Training was done for 120,000 iterations. The batch size was set to 1. Initial learning rate was set to 0.0001, with a 10% drop at $80,000$th and $100,000$th iterations. For the first $15,000$ iterations, the backbone network was kept frozen. We used Stochastic Gradient Descent as an optimizer. The current implementation was heavily adapted from the maskrcnn-benchmark [24] repository and used Pytorch [27] deep learning framework.

For the proposed Context-Aware Regional Attention Network (CARAN), the weights of the fusion module and the rest of the parameters were trained using different subsets of training data. Specifically, we randomly set aside 10% of the training data to train the fusion weights, while the rest of the parameters were kept frozen; then the fusion weights were frozen, while the other parameters were trained using the rest 90% training data.

4.3. Experimental Results

Extensive experiments have been conducted on both Group Affect 2.0 and GroupEmoW datasets demonstrating the effectiveness of the proposed *CARAN* method. Furthermore, to get an insider's view of the proposed network, ablation studies were performed on various components by constructing and evaluating four baseline methods, where findings on the property of each component support the performance improvement the proposed model achieved.

BL_S is the first baseline, which only utilizes the scene stream for classification. Technically, it is very similar to Resnet-50-FPN. Their difference is that the input image is not resized at the input layer, but after the backbone network through $RoIAlign(\cdot)$. BL_{SF} is the second baseline with both face and scene streams. Mean pooling is used when combining the features of multiple RoIs; and average fusion is employed while combining the predictions from the two streams. BL_{SFO} is the third baseline with all three streams, where the feature/score fusion strategies are employed as those in BL_{SF}. *Regional Attention Network (RAN)* is the fourth baseline. Different to the proposed CARAN model, average fusion is employed in *RAN* to combine predictions from multiple streams. All the baseline methods have the same experimental setting as the proposed *CARAN* method. *Note that, all the baseline methods are developed in this work and present part of the contributions.*

Table 3. Comparison of the overall *CARAN* model vs different streams in terms of confusion matrix and recognition accuracy.

		Overall			Scene			Face			Object		
		neu	pos	neg	neu	pos	neg	neu	pos	neg	neu	pos	neg
Group Affect 2.0		acc=**79.13**			acc=76.32			acc=73.29			Acc=77.75		
	neu	75.00	8.48	16.52	68.64	10.60	20.76	69.81	9.43	20.76	**75.80**	10.10	14.18
	pos	9.10	**86.89**	4.01	9.62	85.63	4.75	11.28	83.92	4.81	10.19	85.75	4.06
	neg	18.52	8.77	**72.71**	19.03	10.32	71.65	28.11	09.83	62.06	21.85	9.59	68.56
GroupEmoW		acc=**90.18**			acc=88.74			acc=87.63			acc=89.11		
	neu	**85.05**	7.27	7.68	82.42	6.87	10.71	81.21	8.99	9.80	83.84	6.67	9.49
	pos	3.62	**94.95**	1.43	4.37	93.59	2.03	4.14	94.05	1.81	3.84	93.97	2.19
	neg	9.87	1.39	**88.50**	10.10	1.39	88.50	12.31	2.56	85.13	10.34	1.97	87.69

4.3.1 Experiments on Group Affect Database 2.0

Experimental results in terms of overall recognition accuracy are summarized in Table 4.2 for Group Affect database 2.0. With the shared backbone and RoI aligning, even the first baseline BL_S boosts performance significantly over the Inception-Img method [8] provided by the organizers of the dataset. After adding one more stream or module, there is a performance gain, which asserts the significance of using multiple streams of information from the same backbone network and the effectiveness of the proposed attention module and context-aware fusion module. Furthermore, the proposed *CARAN* performs best compared to all baseline models and also achieves the second best among all the methods in comparison on both "Validation" and "Test" subsets. Note that the winner of the EmotiW2018 GER sub-challenge [13] utilized additional skeleton information, which will be considered in our future work.

4.3.2 Experiments on GroupEmoW Dataset

Table 4.3 compares the proposed *CARAN* model with the baseline models and the state-of-the-art methods for GroupEmoW dataset. It can be observed that the proposed *CARAN* achieves the best result in terms of overall recognition accuracy among all methods compared with. Consistent to the observation on the Group Affect Database 2.0, the performance boosts with inclusion of additional information as well as the proposed attention module and the context-aware module.

4.3.3 Ablation Study on Different Streams

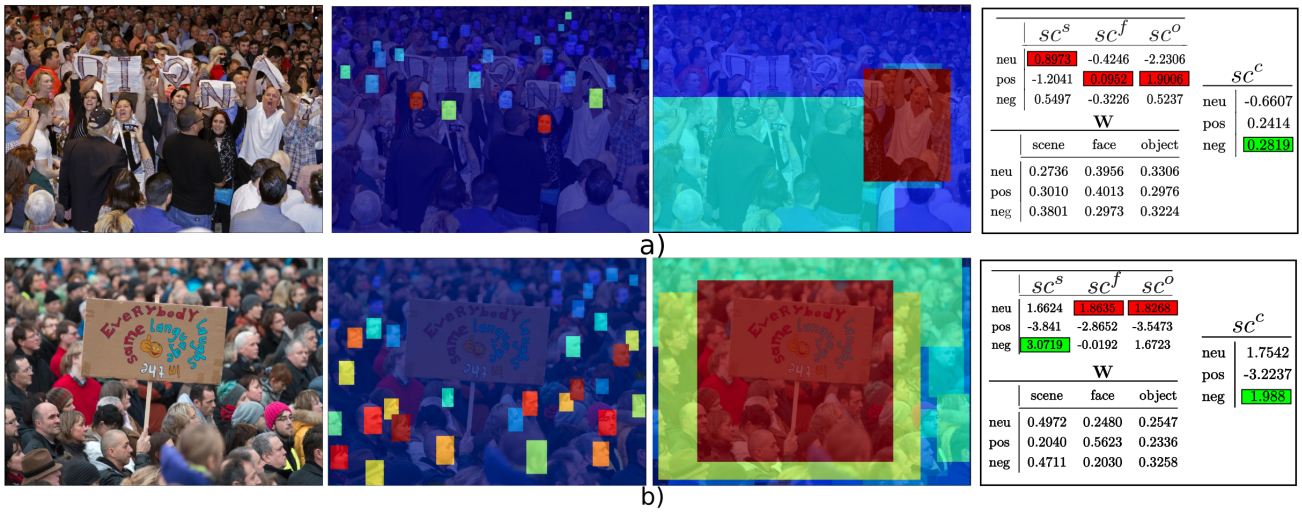
In order to study how different streams contribute to the overall model, the recognition accuracy and confusion matrix are reported in Table 4.3 for each individual stream and the overall model, respectively. Not surprisingly, the overall model outperforms all individual streams in terms of the overall accuracy and individual emotion classes. As shown in Table 4.3, the object stream consistently performs the best among all individual streams for recognizing "Neutral" emotion; while the scene stream beats other streams for recognizing "Negative" emotion. These findings demonstrate that the presence of objects in the scene helps GER especially for "Neutral" and "Negative" emotions, which are difficult to recognize from faces.

4.3.4 Visualization Study for the Attention Module and the Context-aware Fusion Module

We performed a visualization study to understand the proposed attention module and the context-aware fusion module. As illustrated in Fig. 5, each row corresponds to a stream with the highest weight in context-aware fusion and gives three examples with "Positive", "Negative", and "Neutral" emotion, respectively. For each example, the three images show the original image, the attention heatmap of the face stream, and the attention heatmap of the object stream, respectively. On the heatmap, RoIs with hotter color have more attention; and for overlapped RoIs, we choose the maximum among them for displaying.

Visualization of attention revealed important properties aligned with the performance improvement from the proposed attention module. For the face stream, more attention is usually drawn on the faces with bigger size, closer to the camera, or closer to the image center, which are often considered more important from the photographer's view. For example, in Fig. 5 b), the two faces, which are bigger and in front of others, are given more attention; whereas the faces in the background have less attention. Furthermore, we have observed that more attention is given to the person showing stronger deterministic emotion in the group. For instance, in Fig. 5 a), the smiling lady shows explicit signs of emotion and is thus, drawn more attention. For the object stream, we have observed that objects carrying emotional cues, e.g., banners and festoons, and persons with emotional posture and gesture such as shaking or raising hands would receive more attention. For example, the young girl with joyous hand gesture in Fig. 5 d) and the ladies' yoga gestures in Fig. 5 f) get the most attention; whereas banners in Fig. 5 e) has the most attention.

Moreover, to understand what kind of images have higher weight on scene, face, and object streams, we looked



	sc^s	sc^f	sc^o
neu	0.8973	-0.4246	-2.2306
pos	-1.2041	0.0952	1.9006
neg	0.5497	-0.3226	0.5237

	sc^c
neu	-0.6607
pos	0.2414
neg	0.2819

W

	scene	face	object
neu	0.2736	0.3956	0.3306
pos	0.3010	0.4013	0.2976
neg	0.3801	0.2973	0.3224

	sc^s	sc^f	sc^o
neu	1.6624	1.8635	1.8268
pos	-3.841	-2.8652	-3.5473
neg	3.0719	-0.0192	1.6723

	sc^c
neu	1.7542
pos	-3.2237
neg	1.988

W

	scene	face	object
neu	0.4972	0.2480	0.2547
pos	0.2040	0.5623	0.2336
neg	0.4711	0.2030	0.3258

Figure 6. A case study of the proposed context-aware fusion module. For each row, the first three columns show the original image, the attention heatmaps of face and object streams, respectively, and the last column gives the classification scores from the three streams and the final fusion score and also the fusion weight matrix. Both examples have the ground truth emotion labels of "Negative". The scores in red background represent false predictions and the scores in green background represent correct predictions.

at the images where each stream has its highest weight. As illustrated in the first row of Fig. 5, the face stream often has the highest weight for the images where there are clear emotional signals from the faces, or there are fewer objects. The object stream has more confidence where there are indicator objects, e.g., happy hand gesture in Fig. 5 d) or banners in Fig. 5 e). In addition, the object proposals also include faces. The scene stream has highest weight when the scene as a whole becomes more important such as uniforms and background in Fig. 5 g) and i), or the indicator regions are not included in the VQA dataset and hence, not identified as object RoIs, e.g., fire and smoke in Fig. 5 h). This visualization study clearly demonstrates that the proposed context-aware fusion module is capable of adapting weights of different streams to image content automatically.

4.3.5 A Case Study of the Context-aware Fusion

In order to further analyze the effectiveness of the context-aware fusion scheme, we performed a case study by looking at specific examples. As shown in Fig. 6, each row presents an example, where the first three columns show the original image, the attention heatmaps of face and object streams, respectively, and the last column gives the classification scores from the three individual streams and the final fusion score as well as the fusion weight matrix. Both examples have the ground truth labels of "Negative". Note that the scores in red background represent false predictions and the scores in green background represent correct predictions.

All three streams gave wrong predictions in Fig. 6 a), and only the scene stream correctly predicts the group emotion as "Negative" in Fig. 6 b). As a result, *both the majority voting fusion strategy and the average fusion strategy will fail for both examples.* However, the proposed context-aware fusion module yields the correct results. A closer look at the weight matrix $\mathbf{W}$ reveals the answer using Fig. 6 a) as an example. For both the scene and object streams, the lowest weights are assigned to the wrong predicted state, i.e., "Neutral" from the scene stream and "Positive" from the object stream, lowering confidence of the wrong prediction; on the contrary, higher weights are assigned to the correct state, i.e., "Negative", boosting the confidence on the correct prediction. In addition, the lowest weight is assigned to "Negative" for the face stream, subsiding the incorrect assessment. Similar observations can be found in Fig. 6 b).

Therefore, the case study has well demonstrated the effectiveness of the proposed context-aware fusion in assessing the confidence of each stream towards final prediction.

5. CONCLUSION

In this work, we have addressed some pressing concerns on the GER task. Specifically, we proposed a regional attention mechanism to emphasize the most important RoIs and a context-aware fusion scheme to automatically adapt the fusion weights to image content. In addition, a shared backbone FPN network for all streams helps to reduce computation and memory overhead. The proposed CARAN model achieves performance comparable to the state-of-the-art on two publicly available GER datasets. Furthermore, the case study and the visualization study have well demonstrated the effectiveness of the proposed attention mechanism and the context-aware fusion scheme in strengthening and highlighting critical information. In the future, we plan to extend the framework to include more information and apply it to more applications such as sentiment analysis, where modeling various information is desired.

References

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] C. Benitez-Quiroz, Y. Wang, and A. Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999. IEEE, 2017.

[3] J. Cai, Z. Meng, A. Khan, Z. Li, J. O'Reilly, and Y. Tong. Island loss for learning discriminative features in facial expression recognition. In *FG*, pages 302–309. IEEE, 2019.

[4] C. Chen, Z. Wu, and Y.-G. Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 127–131, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *ICMI*, pages 524–528. ACM, 2017.

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

[8] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 653–656, 2018.

[9] H. Ding, S. Zhou, and R. Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *FG*, pages 118–126. IEEE, 2017.

[10] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *ICMI*, pages 445–450, 2016.

[11] X. Guo, L. Polanía, and K. Barner. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *ICMI*, pages 603–608. ACM, 2017.

[12] X. Guo, L. Polania, B. Zhu, C. Boncelet, and K. Barner. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2921–2930, 2020.

[13] X. Guo, B. Zhu, L. Polanía, C. Boncelet, and K. Barner. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *ICMI*, pages 635–639. ACM, 2018.

[14] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli. An attention model for group-level emotion recognition. In *ICMI*, pages 611–615. ACM, 2018.

[15] S. Han, Z. Meng, A. S. Khan, and Y. Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *NIPS*, pages 109–117, 2016.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] A. Khan, Z. Li, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong. Group-level emotion recognition using deep models with a four-stream hybrid network. In *ICMI*, pages 623–629. ACM, 2018.

[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*, pages 1667–1675, 2017.

[20] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019.

[21] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016.

[22] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812, 2014.

[23] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[24] F. Massa and R. Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. `https://github.com/facebookresearch/maskrcnn-benchmark`, 2018. Accessed: [09/16/2019].

[25] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong. Identity-aware convolutional neural network for facial expression recognition. In *FG*, pages 558–565. IEEE, 2017.

[26] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE T-AC*, 2011.

[27] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[28] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[29] K. Schindler, L. Van Gool, and B. De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246, 2008.

[30] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *ICMI*, pages 549–552. ACM, 2017.

[31] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[32] K. Wang, X. Zeng, J. Yang, D. Meng, K. Zhang, X. Peng, and Y. Qiao. Cascade attention networks for group emotion recognition with face, body and image cues. In *ICMI*, pages 640–645. ACM, 2018.

[33] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[34] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018.

[35] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.