EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle

Trisha Mittal[1], Pooja Guhan[1], Uttaran Bhattacharya[1], Rohan Chandra[1], Aniket Bera[1], Dinesh Manocha[1,2]
[1]Department of Computer Science, University of Maryland, College Park, USA
[2]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
{trisha, pguhan, uttaranb, rohan, ab, dm}@cs.umd.edu
Project URL: https://gamma.umd.edu/emoticon

Abstract

We present EmotiCon, a learning-based algorithm for context-aware perceived human emotion recognition from videos and images. Motivated by Frege's Context Principle from psychology, our approach combines three interpretations of context for emotion recognition. Our first interpretation is based on using multiple modalities (e.g. faces and gaits) for emotion recognition. For the second interpretation, we gather semantic context from the input image and use a self-attention-based CNN to encode this information. Finally, we use depth maps to model the third interpretation related to socio-dynamic interactions and proximity among agents. We demonstrate the efficiency of our network through experiments on EMOTIC, a benchmark dataset. We report an Average Precision (AP) score of 35.48 *across* 26 *classes, which is an improvement of* 7-8 *over prior methods. We also introduce a new dataset, GroupWalk, which is a collection of videos captured in multiple real-world settings of people walking. We report an AP of* 65.83 *across* 4 *categories on GroupWalk, which is also an improvement over prior methods.*

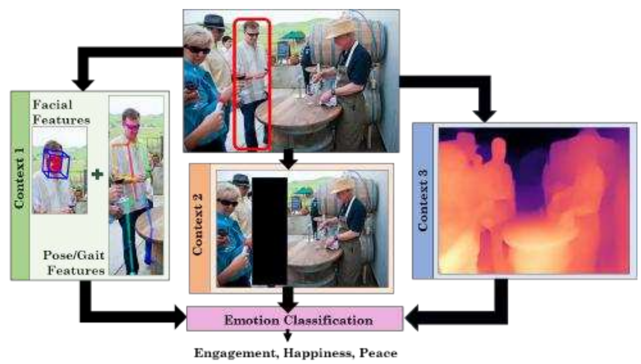

Figure 1: **Context-Aware Multimodal Emotion Recognition:** We use three interpretations of context to perform perceived emotion recognition. We use multiple modalities (Context 1) of faces and gaits, background visual information (Context 2) and socio-dynamic inter-agent interactions (Context 3) to infer the perceived emotion. EmotiCon outperforms prior context-aware emotion recognition methods. Above is an input sample from the EMOTIC dataset.

1. Introduction

Perceiving the emotions of people around us is vital in everyday life. Humans often alter their behavior while interacting with others based on their perceived emotions. In particular, automatic emotion recognition has been used for different applications, including human-computer interaction [13], surveillance [12], robotics, games, entertainment, and more. Emotions are modeled as either discrete categories or as points in a continuous space of affective dimensions [16]. In the continuous space, emotions are treated as points in a 3D space of valence, arousal, and dominance. In this work, our focus is on recognizing perceived human emotion rather than the actual emotional state of a person in the discrete emotion space.

Initial works in emotion recognition have been mostly unimodal [46, 1, 47, 44] approaches. The unique modality may correspond to facial expressions, voice, text, body posture, gaits, or physiological signals. This was followed by multimodal emotion recognition [49, 21, 50], where various combinations of modalities were used and combined in various manners to infer emotions.

Although such modalities or cues extracted from a person can provide us with information regarding the perceived emotion, context also plays a very crucial role in the understanding of the perceived emotion. Frege's context principle [45] urges not asking for the meaning of a word in isolation and instead of finding the meaning in the context of a sentence. We use this notion behind the context principle in psychology for emotion recognition. 'Context' has been interpreted in multiple ways by researchers in psychology, including:

(a) *Context 1 (Multiple Modalities):* Incorporating cues from different modalities was one of the initial definitions of context. This domain is also known as Multimodal Emotion Recognition. Combining modalities provides complementary information, which leads to better inference and also performs better on in-the-wild datasets.

(b) *Context 2 (Background Context):* Semantic understanding of the scene from visual cues in the image helps in getting insights about the agent's surroundings and activity, both of which can affect the perceived emotional state of the agent.

(c) *Context 3 (Socio-Dynamic Inter-Agent Interactions):* Researchers in psychology suggest that the presence or absence of other agents affects the perceived emotional state of an agent. When other agents share an identity or are known to the agent, they often coordinate their behaviors. This varies when other agents are strangers. Such interactions and proximity to other agents have been less explored for perceived emotion recognition.

One of our goals is to make Emotion Recognition systems work for real-life scenarios. This implies using modalities that do not require sophisticated equipment to be captured and are readily available. Psychology researchers [3] have conducted experiments by mixing faces and body features corresponding to different emotions and found that participants guessed the emotions that matched the body features. This is also because of the ease of "mocking" one's facial expressions. Subsequently, researchers [25, 38] found the combination of faces and body features to be a reliable measure of inferring human emotion. As a result, it would be useful to combine such face and body features for context-based emotion recognition.

Main Contributions: We propose EmotiCon, a context-aware emotion recognition model. The input to EmotiCon is images/video frames, and the output is a multi-label emotion classification. The novel components of our work include:

1. We present a context-aware multimodal emotion recognition algorithm called EmotiCon. Consistent with Ferge's Context principle, in this work, we try to incorporate three interpretations of context to perform emotion recognition from videos and images.
2. We also present a new approach to modeling the socio-dynamic interactions between agents using a depth-based CNN. We compute a depth map of the image and feed that to the network to learn about the proximity of agents to each other.
3. Though extendable to any number of modalities, we release a new dataset GroupWalk for emotion recognition. To the best of our knowledge, there exist very few datasets captured in uncontrolled settings with both faces and gaits that have emotion label annotations. To enable research in this domain, we make GroupWalk publicly available with emotion annotations. GroupWalk is a collection of 45 videos captured in multiple real-world settings of people walking in dense crowd settings. The videos have about 3544 agents annotated with their emotion labels.

We compare our work with prior methods by testing our performance on EMOTIC [28], a benchmark dataset for context-aware emotion recognition. We report an improved AP score of 35.48 on EMOTIC, which is an improvement of $7-8$ over prior methods [27, 30, 58]. We also report AP scores of our approach and prior methods on the new dataset, GroupWalk. We perform ablation experiments on both datasets, to justify the need for the three components of EmotiCon. As per the annotations provided in EMOTIC, we perform a multi-label classification over 26 discrete emotion labels. On GroupWalk too, we perform a multi-label classification over 4 discrete emotions (anger, happy, neutral, sad).

2. Related Work

In this section, we give a brief overview of previous works on unimodal and multimodal emotion recognition, context-aware emotion recognition, and existing context-aware datasets.

2.1. Uni/Multimodal Emotion Recognition

Prior works in emotion recognition from handcrafted features [48, 60] or deep learning networks [32, 17, 31] have used single modalities like facial expressions [46, 1], voice, and speech expressions [47], body gestures [42], gaits [44], and physiological signals such as respiratory and heart cues [26]. There has been a shift in the paradigm, where researchers have tried to fuse multiple modalities to perform emotion recognition, also known as Multimodal Emotion Recognition. Fusion methods like early fusion [49], late fusion [21], and hybrid fusion [50] have been explored for emotion recognition from multiple modalities. Multimodal emotion recognition has been motivated by research in psychology and also helped in improving accuracy on in-the-wild emotion recognition datasets like IEMOCAP [9] and CMU-MOSEI [57].

2.2. Context-Aware Emotion Recognition in Psychology Research

Though introduced in the domain of philosophy of language, Frege [45] proposed that words should never be seen in isolation but in the context of their proposition. Researchers in psychology [6, 29, 37] also agree that just like most psychological processes, emotional processes cannot be interpreted without context. They suggest that context often produces emotion and also shapes how emotion is perceived. Emotion literature that addresses context [2, 5, 39] suggests several broad categories of contextual features: person, situation, and context. Martinez et al. [36] conduct experiments about the necessity of context and found that even when the participants' faces and bodies were masked in silent videos, viewers were able to infer the affect successfully. Greenway et al. [20] organize these contextual features in three levels, ranging from micro-level (person) to macro-level (cultural). In level 2 (situational), they include factors like the presence and closeness

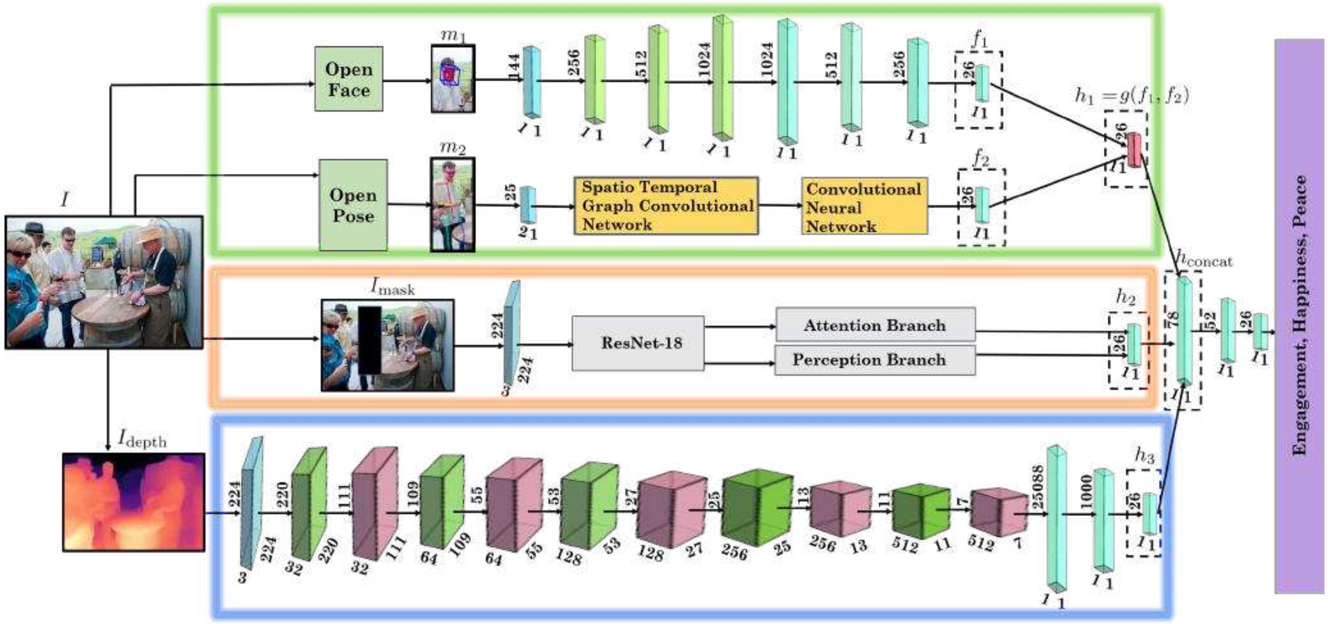

Figure 2: **EmotiCon:** We use three interpretations of context. We first extract features for the two modalities to obtain f_1 and f_2 and inputs I_{mask} and I_{depth} from the raw input image, I. These are then passed through the respective neural networks to obtain h_1, h_2 and h_3. To obtain h_1, we use a multiplicative fusion layer (red color) to fuse inputs from both modalities, faces, and gaits. h_1, h_2 and h_3 are then concatenated to obtain h_{concat}.

of other agents. Research shows that the simple presence of another person elicits more expression of emotion than situations where people are alone [54, 24]. These expressions are more amplified when people know each other and are not strangers [24].

2.3. Context-Aware Emotion Recognition

Recent works in context-aware emotion recognition are based on deep-learning network architectures. Kosti et al. [27] and Lee et al. [30] present two recent advances in context-aware emotion recognition and they propose similar architectures. Both of them have two-stream architectures followed by a fusion network. One stream focuses on a modality (face for [30] and body for [27]) and the other focuses on capturing context. Lee et al. [30] consider everything other than the face as context, and hence mask the face from the image to feed to the context stream. On the other hand, [30] uses a Region Proposal Network (RPN) to extract context elements from the image. These elements become the nodes of an affective graph, which is fed into a Graph Convolution Network (GCN) to encode context. Another problem that has been looked into is group emotion recognition [19, 53]. The objective here is to label the emotion of the entire set of people in the frame under the assumption that they all share some social identity.

2.4. Context-Aware Emotion Recognition Datasets

Most of the emotion recognition datasets in the past have either only focused on a single modality, e.g., faces or body features, or have been collected in controlled settings. For example, the GENKI database [52] and the UCD-SEE dataset [51] are datasets that focus primarily on the facial expressions collected in lab settings. The Emotion Recognition in the Wild (EmotiW) challenges [14] host three databases: AFEW dataset [15] (collected from TV shows and movies), SFEW (a subset of AFEW with only face frames annotated), and HAPPEI database, which focuses on the problem of group-level emotion estimation. Some of the recent works have realized the potential of using context for emotion recognition and highlighted the lack of such datasets. Context-Aware Emotion Recognition (CAER) dataset [58] is a collection of video-clips from TV shows with 7 discrete emotion annotations. EMOTIC dataset [27] is a collection of images from datasets like MSCOCO [34] and ADE20K [61] along with images downloaded from web searches. The dataset is a collection of $23,571$ images, with about $34,320$ people annotated for 26 discrete emotion classes. We have summarised and compared all these datasets in Table 1.

3. Our Approach: EmotiCon

In this section, we give an overview of the approach in Section 3.1 and motivate the three context interpretations in Section 3.2, 3.3, and 3.4.

3.1. Notation and Overview

We present an overview of our context-aware multimodal emotion recognition model, EmotiCon, in Figure 2. Our input consists of an RGB image, I. We process I to

Data type	Dataset	Dataset Size	Agents Annotated	Setting	Emotion Labels	Context
Images	EMOTIC [27]	18,316 images	34,320	Web	26 Categories	Yes
	AffectNet [41]	450,000 images	450,000	Web	8 Categories	No
	CAER-S [30]	70,000 images	70,000	TV Shows	7 Categories	Yes
Videos	AFEW [15]	1,809 clips	1,809	Movie	7 Categories	No
	CAER [30]	13,201 clips	13,201	TV Show	7 Categories	Yes
	IEMOCAP [9]	12 hrs	-	TV Show	4 Categories	Yes
	GroupWalk	45 clips(10 mins each)	3544	Real Settings	4 Categories	Yes

Table 1: **Context-Aware Emotion Recognition Dataset Analysis:** We compare GroupWalk with existing emotion recognition datasets such as EMOTIC [27], AffectNet [41], CAER and CAER-S [30], and AFEW [15].

generate the input data for each network corresponding to the three contexts. The network for Context 1 consists of n streams corresponding to n distinct modalities denoted as $m_1, m_2, \ldots, m_n$. Each distinct layer outputs a feature vector, f_i. The n feature vectors $f_1, f_2, \ldots, f_n$ are combined via multiplicative fusion [40] to obtain a feature encoding, $h_1 = g(f_1, f_2, \ldots, f_n)$, where $g(\cdot)$ corresponds to the multiplicative fusion function. Similarly, h_2, and h_3 are computed through the networks corresponding to the second and third Contexts. h_1, h_2, and h_3 are concatenated to perform multi-label emotion classification.

3.2. Context 1: Multiple Modalities

In real life, people appear in a multi-sensory context that includes a voice, a body, and a face; these aspects are also perceived as a whole. Combining more than one modality to infer emotion is beneficial because cues from different modalities can complement each other. They also seem to perform better on in-the-wild datasets [40] than other unimodal approaches. Our approach is extendable to any number of modalities available. To validate this claim, other than EMOTIC and GroupWalk, which have two modalities, faces, and gaits, we also show results on the IEMOCAP dataset which face, text, and speech as three modalities. From the input image I, we obtain $m_1, m_2, \ldots, m_n$ using processing steps as explained in Section 4.1. These inputs are then passed through their respective neural network architectures to obtain $f_1, f_2, \ldots, f_n$. To make our algorithm robust to sensor noise and averse to noisy signals, we combine these features multiplicatively to obtain h_1. As shown in previous research [35, 40], multiplicative fusion learns to emphasize reliable modalities and to rely less on other modalities. To train this, we use the modified loss function proposed previously [40] defined as:

$$L_{\text{multiplicative}} = -\sum_{i=1}^{n} (p_i^e)^{\frac{\beta}{n-1}} \log p_i^e \tag{1}$$

where n is the total number of modalities being considered, and p_i^e is the prediction for emotion class, e, given by the network for the i^{th} modality.

3.3. Context 2: Situational/Background Context

Our goal is to identify semantic context from images and videos to perform perceived emotion recognition. Semantic context includes the understanding of objects –excluding the primary agent– present in the scene, their spatial extents, keywords, and the activity being performed. For instance, in Figure 1, the input image consists of a group of people gathered around with drinks on a bright sunny day. The “bright sunny day”, “drink glasses”, “hats” and “green meadows” constitute semantic components and may affect judgement of one’s perceived emotion.

Motivated by multiple approaches in the computer vision literature [59, 18] surrounding semantic scene understanding, we use an attention mechanism to train a model to focus on different aspects of an image while *masking* the primary agent, to extract the semantic components of the scene. The mask, $I_{\text{mask}} \in \mathbb{R}^{224\times224}$, for an input image I is given as

$$I_{\text{mask}} = \begin{cases} I(i,j) & \text{if } I(i,j) \notin \text{bbox}_{\text{agent}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\text{bbox}_{\text{agent}}$ denotes the bounding box of the agent in the scene.

3.4. Context 3: Inter-Agent Interactions/Socio-Dynamic Context

When an agent is surrounded by other agents, their perceived emotions change. When other agents share an identity or are known to the agent, they often coordinate their behaviors. This varies when other agents are strangers. Such interactions and proximity can help us infer the emotion of agents better.

Prior experimental research has used walking speed, distance, and proximity features to model socio-dynamic interactions between agents to interpret their personality traits. Some of these algorithms, like the social force model [23], are based on the assumption that pedestrians are subject to attractive or repulsive forces that drive their dynamics. Non-linear models like RVO [56] aim to model collision avoidance among individuals while walking to their individual goals. But, both of these methods do not capture cohesiveness in a group.

We propose an approach to model these socio-dynamic interactions by computing proximity features using depth maps. The depth map, $I_{\text{depth}} \in \mathbb{R}^{224\times224}$, corresponding to input image, I, is represented through a 2D matrix where,

$$I_{\text{depth}}(i,j) = d(I(i,j), c) \tag{3}$$

$d(I(i,j), c)$ represents the distance of the pixel at the i^{th} row and j^{th} column from the camera center, c. We pass I_{depth} as input depth maps through a CNN and obtain h_3.

In addition to depth map-based representation, we also use Graph Convolutional Networks (GCNs) to model the proximity-based socio-dynamic interactions between agents. GCNs have been used to model similar interactions in traffic networks [22] and activity recognition [55]. The input to a GCN network consists of the spatial coordinates of all agents, denoted by $X \in \mathbb{R}^{n\times 2}$, where n represents the number of agents in the image, as well as the unweighted adjacency matrix, $A \in \mathbb{R}^{n\times n}$, of the agents, which is defined as follows,

$$A(i,j) = \begin{cases} e^{-d(v_i,v_j)} & \text{if } d(v_i,v_j) < \mu, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The function $f = e^{-d(v_i,v_j)}$ [7] denotes the interactions between any two agents.

4. Network Architecture and Implementation Details

In this section, we elaborate on the implementation and network architectures of EmotiCon. The data preprocessing for the streams of EmotiCon are presented in 4.1. We include details about the network architectures of context 1, context 2, and context 3 in Section 4.2. We explain the early fusion technique we use to fuse the features from the three context streams to infer emotion and the loss function used for training the multi-label classification problem.

4.1. Data Processing

Context1: We use OpenFace [4] to extract a 144-dimensional face modality vector, $m_1 \in \mathbb{R}^{144}$ obtained through multiple facial landmarks. We compute the 2D gait modality vectors, $m_2 \in \mathbb{R}^{25\times 2}$ using OpenPose [10] to extract 25-coordinates from the input image I. For each coordinate, we record the x and y pixel values.

Context2: We use RobustTP [11], which is a pedestrian tracking method to compute the bounding boxes for all agents in the scene. These bounding boxes are used to compute I_{mask} according to Equation 2.

Context3: We use Megadepth [33] to extract the depth maps from the input image I. The depth map, I_{depth}, is computed using Equation 3.

4.2. Network Architecture

Context1: Given a face vector, m_1, we use three 1D convolutions (depicted in light green color in Figure 2) with batch normalization and ReLU non-linearity. This is followed by a max pool operation and three fully-connected layers (cyan color in Figure 2) with batch normalization and ReLU. For m_2, we use the ST-GCN architecture proposed by [8], which is currently the SOTA network for emotion classification using gaits. Their method was originally designed to deal with 2D pose information for 16 body joints. We modify their setup for 2D pose inputs for 25 joints. We show the different layers and hyper-parameters used in Figure 2. The two networks give us f_1 and f_2, which are then multiplicatively fused (depicted in red color in Figure 2) to generate h_1.

Context 2: For learning the semantic context of the input image I, we use the Attention Branch Network (ABN) [18] on the masked image I_{mask}. ABN contains an attention branch which focuses on attention maps to recognize and localize important regions in an image. It outputs these potentially important locations in the form of h_2.

Context 3: We perform two experiments using both depth map and a GCN. For depth-based network, we compute the depth map, I_{depth} and pass it through a CNN. The CNN is composed of 5 alternating 2D convolutional layers (depicted in dark green color in Figure 2) and max pooling layers (magenta color in Figure 2). This is followed by two fully connected layers of dimensions 1000 and 26 (cyan color in Figure 2).

For the graph-based network, we use two graph convolutional layers followed by two linear layers of dimension 100 and 26.

Fusing Context Interpretations: To fuse the feature vectors from the three context interpretations, we use an early fusion technique. We concatenate the feature vectors before making any individual emotion inferences.

$$h_{\text{concat}} = [h_1, h_2, h_3]$$

We use two fully connected layers of dimensions 52 and 26, followed by a softmax layer. This output is used for computing the loss and the error, and then back-propagating the error back to the network.

Loss Function: Our classification problem is a multi-label classification problem where we assign one or more than one emotion label to an input image or video. To train this network, we use the multi-label soft margin loss function and denote it by $L_{\text{classification}}$. The loss function optimizes a multi-label one-versus-all loss based on max-entropy between the input x and output y.

So, we combine the two loss functions, $L_{\text{multiplicative}}$ (from Eq. 1) and $L_{\text{classification}}$ to train EmotiCon.

$$L_{\text{total}} = \lambda_1 L_{\text{multiplicative}} + \lambda_2 L_{\text{classification}} \tag{5}$$

5. Datasets

In Section 5.1, we give details about the benchmark dataset for context-aware emotion recognition, EMOTIC. We present details about the new dataset, GroupWalk and also perform a comparison with other existing datasets in Section 5.2. Like summarised in Table 1, there are a lot more datasets for emotion recognition, but they do not have any context available. Though our approach will work on these datasets, we do not expect any significant improvement over the SOTA on these datasets. Just to reinforce this, we did run our method on IEMOCAP [9], which has limited context information, and summarise our results in Appendix B.

5.1. EMOTIC Dataset

The EMOTIC dataset contains 23,571 images of 34,320 annotated people in unconstrained environments. The annotations consist of the apparent emotional states of the people in the images. Each person is annotated for 26 discrete categories, with multiple labels assigned to each image.

5.2. GroupWalk Dataset

5.2.1 Annotation

GroupWalk consists of 45 videos that were captured using stationary cameras in 8 real-world setting including a hospital entrance, an institutional building, a bus stop, a train station, and a marketplace, a tourist attraction, a shopping place and more. The annotators annotated agents with clearly visible faces and gaits across all videos. 10 annotators annotated a total of 3544 agents. The annotations consist of the following emotion labels– Angry, Happy, Neutral, and Sad. Efforts to build on this dataset are still ongoing. The dataset collected and annotated so far can be found at the Project webpage. To prepare train and test splits for the dataset, we randomly selected 36 videos for the training and 9 videos for testing.

While perceived emotions are essential, other affects such as dominance and friendliness are important for carrying out joint and/or group tasks. Thus, we additionally label each agent for dominance and friendliness. More details about the annotation process, labelers and labels processing are presented in Appendix A.

6. Experiments and Results

In this section, we discuss the experiments conducted for EmotiCon. We present details on hyperparameters and training details in Section 6.1. In section 6.2, we list the prior methods we compare the performance of EmotiCon with. We present an elaborate analysis of both qualitative and quantitative results in Section 6.3. In Section 6.5, we perform experiments to validate the importance of each component of EmotiCon.

6.1. Training Details

For training EmotiCon on the EMOTIC dataset, we use the standard train, val, and test split ratios provided in the dataset. For GroupWalk, we split the dataset into 85% training (85%) and testing (15%) sets. In GroupWalk each sample point is an agent ID; hence the input is all the frames for the agent in the video. To extend EmotiCon on videos, we perform a forward pass for all the frames and take an average of the prediction vector across all the frames and then compute the AP scores and use this for loss calculation and backpropagating the loss. We use a batch size of 32 for EMOTIC and a batchsize of 1 for GroupWalk. We train EmotiCon for 75 epochs. We use the Adam optimizer with a learning rate of 0.0001. All our results were generated on NVIDIA GeForce GTX 1080 Ti GPU. All the code was implemented using PyTorch [43].

6.2. Evaluation Metrics and Methods

We use the standard metric Average Precision (AP) to evaluate all our methods. For both EMOTIC and GroupWalk datasets, we compare our methods with the following SOTA methods.

1. **Kosti et al. [27]** propose a two-stream network followed by a fusion network. The first stream encodes context and then feeds the entire image as an input to the CNN. The second stream is a CNN for extracting body features. The fusion network combines features of the two CNNs and estimates the discrete emotion categories.
2. **Zhang et al. [58]** build an affective graph with nodes as the context elements extracted from the image. To detect the context elements, they use a Region Proposal Network (RPN). This graph is fed into a Graph Convolutional Network (GCN). Another parallel branch in the network encodes the body features using a CNN. The outputs from both the branches are concatenated to infer an emotion label.
3. **Lee et al. [30]** present a network architecture, CAER-Net consisting of two subnetworks, a two-stream encoding network, and an adaptive fusion network. The two-stream encoding network consists of a face stream and a context-stream where facial expression and context (background) are encoded. An adaptive fusion network is used to fuse the two streams.

We use the publicly available implementation for Kosti et al. [27] and train the entire model on GroupWalk. Both Zhang et al. [58] and Lee et al. [30] do not have publicly available implementations. We reproduce the method by Lee et al. [30] to the best of our understanding. For Zhang et al. [58], while we report their performance on the EMOTIC dataset, with limited implementation details, it was difficult to build their model to test their performance on GroupWalk.

6.3. Analysis and Discussion

Comparison with SOTA: We summarize the evaluation of the APs for all the methods on the EMOTIC and GroupWalk datasets in Table 2. For EmotiCon, we report the AP scores for both GCN-based and Depth Map-based implementations of Context 3. On both the EMOTIC and GroupWalk datasets, EmotiCon outperforms the SOTA.

Generalize to more Modalities: A major factor for the success of EmotiCon is its ability to combine different

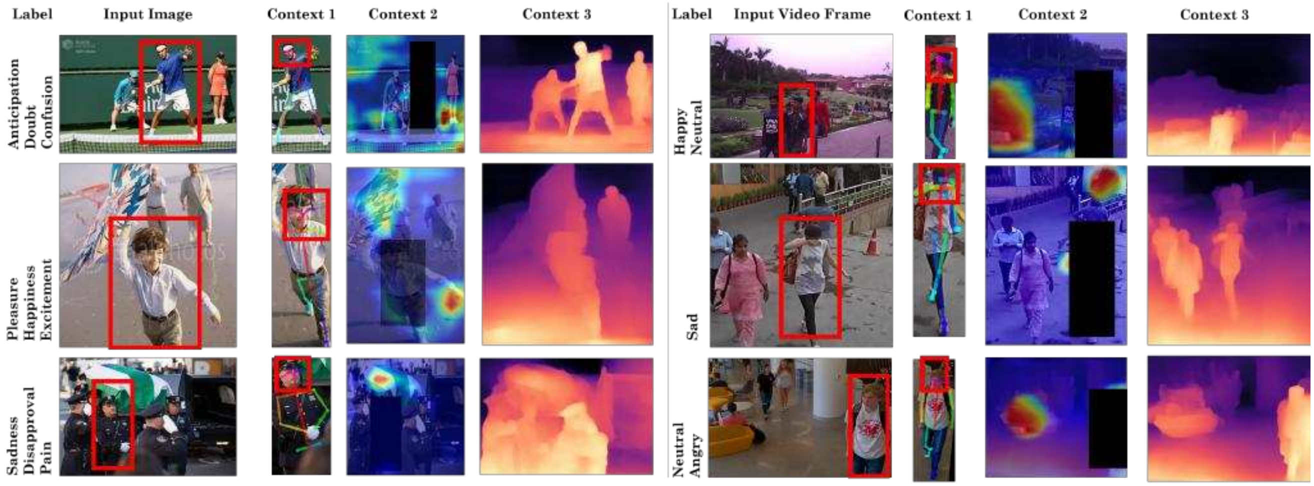

Figure 3: **Qualitative Results:** We show the classification results on three examples, each from the EMOTIC dataset (left) and GroupWalk Dataset (right), respectively. In the top row example (left) and middle row example (right), the depth map clearly marks the tennis player about to swing to convey anticipation, and the woman coming from the hospital to convey sadness, respectively. In the bottom row (left) and bottom row (middle) examples, the semantic context of the coffin and the child's kite is clearly identified to convey sadness and pleasure, respectively.

(a) AP Scores for EMOTIC Dataset.

Labels	Kosti et al.[27]	Zhang et al.[58]	Lee et al.[30]	**EmotiCon**	
				GCN-Based	Depth-Based
Affection	27.85	46.89	19.9	36.78	**45.23**
Anger	09.49	10.87	11.5	14.92	**15.46**
Annoyance	14.06	11.23	16.4	18.45	**21.92**
Anticipation	58.64	62.64	53.05	68.12	**72.12**
Aversion	07.48	5.93	16.2	16.48	**17.81**
Confidence	78.35	72.49	32.34	59.23	**68.65**
Disapproval	14.97	11.28	16.04	**21.21**	19.82
Disconnection	21.32	26.91	22.80	25.17	**43.12**
Disquietment	16.89	16.94	17.19	16.41	**18.73**
Doubt/Confusion	29.63	18.68	28.98	33.15	**35.12**
Embarrassment	03.18	1.94	15.68	11.25	**14.37**
Engagement	87.53	88.56	46.58	90.45	**91.12**
Esteem	17.73	13.33	19.26	22.23	**23.62**
Excitement	77.16	71.89	35.26	82.21	**83.26**
Fatigue	09.70	13.26	13.04	**19.15**	16.23
Fear	14.14	4.21	10.41	11.32	**23.65**
Happiness	58.26	73.26	49.36	68.21	**74.71**
Pain	08.94	6.52	10.36	12.54	**13.21**
Peace	21.56	32.85	16.72	**35.14**	34.27
Pleasure	45.46	57.46	19.47	61.34	**65.53**
Sadness	19.66	25.42	11.45	**26.15**	23.41
Sensitivity	09.28	5.99	**10.34**	9.21	8.32
Suffering	18.84	23.39	11.68	22.81	**26.39**
Surprise	**18.81**	9.02	10.92	14.21	17.37
Sympathy	14.71	17.53	17.125	24.63	**34.28**
Yearning	08.34	10.55	9.79	12.23	**14.29**
mAP	27.38	28.42	20.84	32.03	**35.48**

(b) AP Scores for GroupWalk Dataset.

Labels	Kosti et al.[27]	Zhang et al.[58]	Lee et al.[30]	**EmotiCon**	
				GCN-Based	Depth-Based
Anger	58.46	-	42.31	65.13	**69.42**
Happy	69.12	-	56.79	72.46	**73.18**
Neutral	42.27	-	39.24	44.51	**48.51**
Sad	63.83	-	54.33	68.25	**72.24**
mAP	58.42	-	48.21	62.58	**65.83**

Table 2: **Emotion Classification Performance:** We report the AP scores on the EMOTIC and the GroupWalk datasets. EmotiCon outperforms all the three methods for most of the classes and also overall.

modalities effectively via multiplicative fusion. Our approach learns to assign higher weights to more expressive modalities while suppressing weaker ones. For example, in instances where the face may not be visible, EmotiConinfers the emotion from context (See Figure 3, middle row(right)). This is in contrast to Lee et al. [30], which relies on the availability of face data. Consequently, they perform poorly on both the EMOTIC and GroupWalk datasets, as both datasets contain many examples where the face is not visible clearly. To further demonstrate the ability of EmotiCon to generalize to any modality, we additionally report our performance on the IEMOCAP dataset [9] in Appendix B.

GCN versus Depth Maps: GCN-based methods do not perform as well as depth-based but are a close second. This may be due to the fact that, on average most images of the EMOTIC dataset contain 5 agents. GCN-based methods in the literature have been trained on datasets with a lot more number of agents in each image or video. Moreover, with a depth-based approach, EmotiCon leans a 3D aspect of the scene in general and is not limited to inter-agent interactions.

Failure Cases: We show two examples from EMOTIC dataset in Figure 4 where EmotiCon fails to classify correctly. We also show the ground-truth and predicted emotion labels. In the first image, EmotiConis unable to gather any context information. On the other hand, in the second image, there is a lot of context information like the many visual elements in the image and multiple agents. This leads to an incorrect inference of the perceived emotion.

6.4. Qualitative Results

We show qualitative results for three examples, each from both the datasets, respectively, in Figure 3. The first column is the input image marking the primary agents, the second column shows the corresponding extracted face and gait, the third column shows the attention maps learned by the model, and lastly, in the fourth column, we show the

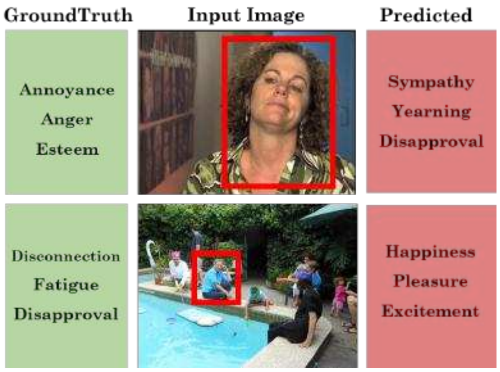

Figure 4: **Misclassification by EmotiCon:** We show two examples where EmotiCon incorrectly classifies the labels. In the first examples, EmotiCon is confused about the prediction due to lack of any context. In the second example, there is a lot of context available, which also becomes confusing.

(a) Ablation Experiments on EMOTIC Dataset.

Labels	Context Interpretations			
	Only 1	Only 1 and 2	Only 1 and 3	**1, 2 and 3**
Affection	29.87	41.83	30.15	45.23
Anger	08.52	11.41	8.36	15.46
Annoyance	09.65	17.37	12.91	21.92
Anticipation	46.23	67.59	60.53	72.12
Aversion	06.27	11.71	09.46	17.81
Confidence	51.92	65.27	59.63	68.85
Disapproval	11.81	17.35	15.41	19.82
Disconnection	31.74	41.46	32.56	43.12
Disquietment	07.57	12.69	12.24	18.73
Doubt/Confusion	21.62	31.28	29.51	35.12
Embarrassment	08.43	10.51	12.25	14.37
Engagement	78.68	84.62	81.51	91.12
Esteem	18.32	18.79	09.42	23.62
Excitement	73.19	80.54	76.14	83.26
Fatigue	06.34	11.95	14.15	16.23
Fear	14.29	21.36	22.29	23.65
Happiness	52.52	69.51	71.51	74.71
Pain	05.75	09.56	11.10	13.21
Peace	13.53	30.72	30.15	34.27
Pleasure	58.26	61.89	59.81	65.53
Sadness	19.94	19.74	22.27	23.41
Sensitivity	03.16	04.11	8.15	8.32
Suffering	15.38	20.92	12.83	26.39
Surprise	05.29	16.45	16.26	17.37
Sympathy	22.38	30.68	22.17	34.28
Yearning	04.94	10.53	9.82	14.29
mAP	24.06	31.53	29.63	35.48

(b) Ablation Experiments on GroupWalk Dataset.

Labels	Context Interpretations			
	Only 1	Only 1 and 2	Only 1 and 3	**1, 2 and 3**
Anger	58.51	63.83	66.15	69.42
Happy	61.24	64.16	68.87	73.18
Neutral	40.36	41.57	44.15	48.51
Sad	62.17	67.22	70.35	72.24
mAP	55.57	59.20	62.38	65.83

Table 3: **Ablation Experiments:** Keeping the Context interpretation 1 throughout, we remove the other two Context interpretations one by one and compare the AP scores for emotion classification on both the datasets.

depth map extracted from the input image.

The heatmaps in the attention maps indicate what the network has learned. In the bottom row (left) and bottom row (middle) examples, the semantic context of the coffin and the child's kite is clearly identified to convey sadness and pleasure, respectively. The depth maps corresponding to the input images capture the idea of proximity and inter-agent interactions. In the top row example (left) and middle row example (right), the depth map clearly marks the tennis player about to swing to convey anticipation, and the woman coming from the hospital to convey sadness, respectively.

6.5. Ablation Experiments

To motivate the importance of Context 2 and Context 3, we run EmotiCon on both EMOTIC and GroupWalk dataset removing the networks corresponding to both contexts, followed by removing either of them one by one. The results of the ablation experiments have been summarized in Table 3. We choose to retain Context 1 in all these runs because it is only Context 1 that is capturing information from the agent itself.

We observe from the qualitative results in Figure 3 that Context 2 seems more expressive in the images of EMOTIC dataset, while Context 3 is more representative in GroupWalk. This is supported by the results reported in Table 3, columns 2 and 3. To understand why this happens, we analyse the two datasets closely. EMOTIC dataset was collected for the task of emotion recognition with context. it is a dataset of pictures collected from multiple datasets and scraped from the Internet. As a result, most of these images have a rich background context. Moreover we also found that more than half the images of EMOTIC contain at most 3 people. These are the reasons we believe that interpretation 2 helps more in EMOTIC than interpretation 3. In the GroupWalk Dataset, the opposite is true. The number of people per frame is much higher. This density gets captured best in interpretation 3 helping the network to make the better inference.

7. Conclusion, Limitations, and Future Work

We present EmotiCon, a context-aware emotion recognition system that borrows and incorporates the context interpretations from psychology. We use multiple modalities (faces and gaits), situational context, and also the socio-dynamic context information. We make an effort to use easily available modalities that can be easily captured or extracted using commodity hardware (e.g., cameras). To foster more research on emotion recognition with naturalistic modalities, we also release a new dataset called GroupWalk. Our model has limitations and often confuses between certain class labels. Further, we currently perform multi-class classification over discrete emotion labels. In the future, we would also like to move towards the continuous model of emotions (Valence, Arousal, and Dominance). As part of future work, we would also explore more such context interpretations to improve the accuracies.

8. Acknowledgements

This research was supported in part by ARO Grants W911NF1910069 and W911NF1910315 and Intel.

References

[1] Kingsley Oryina Akputu, Kah Phooi Seng, and Yun Li Lee. Facial emotion recognition for intelligent tutoring environment. In *IMLCS*, pages 9–13, 2013. 1, 2

[2] Amelia Aldao. The future of emotion regulation research: Capturing context. *Perspectives on Psychological Science*, 8(2):155–172, 2013. 2

[3] H Aviezer, Y Trope, et al. Body cues, not facial expressions, discriminate between intense positive and negative emotions. 2012. 2

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 5

[5] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. 2

[6] Lisa Feldman Barrett, Batja Mesquita, and Eliot R Smith. The context principle. *The mind in context*, 1, 2010. 2

[7] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 5

[8] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. *arXiv preprint arXiv:1910.12906*, 2019. 5

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008. 2, 4, 6, 7, 11

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 5

[11] Rohan Chandra, Uttaran Bhattacharya, Christian Roncal, Aniket Bera, and Dinesh Manocha. Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In *ACM Computer Science in Cars Symposium*, page 2. ACM, 2019. 5

[12] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50:487–503, 06 2008. 1

[13] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *SP Magazine, IEEE*, 18:32 – 80, 02 2001. 1

[14] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 427–432. ACM, 2016. 3

[15] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012. 3, 4

[16] Paul Ekman and Wallace V Friesen. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, 24(3 PT 1):711–724, 1967. 1

[17] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix 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*, pages 5562–5570, 2016. 2

[18] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019. 4, 5

[19] Samanyou Garg. Group emotion recognition using machine learning, 2019. 3

[20] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. Context is everything (in emotion research). *Social and Personality Psychology Compass*, 12(6):e12393, 2018. 2

[21] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007. 1, 2

[22] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019. 5

[23] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 4

[24] Esther Jakobs, Antony SR Manstead, and Agneta H Fischer. Social context effects on facial activity in a negative emotional setting. *Emotion*, 1(1):51, 2001. 3

[25] A Kleinsmith, N Bianchi-Berthouze, et al. Affective body expression perception and recognition: A survey. *IEEE TAC*, 2013. 2

[26] R Benjamin Knapp, Jonghwa Kim, and Elisabeth André. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In *Emotion-oriented systems*, pages 133–159. Springer, 2011. 2

[27] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 3, 4, 6, 7, 11, 12

[28] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[29] Alison Ledgerwood. Evaluations in their social context: Distance regulates consistency and context dependence. *Social and Personality Psychology Compass*, 8(8):436–447, 2014. 2

[30] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. *arXiv preprint arXiv:1908.05913*, 2019. 2, 3, 4, 6, 7, 11

[31] Shan Li, Weihong Deng, and JunPing 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*, pages 2852–2861, 2017. 2

[32] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with

attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 2

[33] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 5

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[35] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018. 4

[36] Aleix M Martinez. Context may reveal how you feel. *Proceedings of the National Academy of Sciences*, 116(15):7169–7171, 2019. 2

[37] James K McNulty and Frank D Fincham. Beyond positive psychology? toward a contextual view of psychological processes and well-being. *American Psychologist*, 67(2):101, 2012. 2

[38] H Meeren, C van Heijnsbergen, et al. Rapid perceptual integration of facial expression and emotional body language. *PNAS*, 2005. 2

[39] Batja Mesquita and Michael Boiger. Emotions in context: A sociodynamic model of emotions. *Emotion Review*, 6(4):298–302, 2014. 2

[40] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, 2019. 4

[41] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019. 4

[42] Costanza Navarretta. Individuality in communicative bodily behaviours. In *Cognitive Behavioural Systems*, pages 417–423. Springer, 2012. 2

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 6

[44] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Uttaran Bhattacharya, Kurt Gray, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, 2019. 1, 2

[45] Michael David Resnik. The context principle in frege's philosophy. *Philosophy and Phenomenological Research*, 27(3):356–365, 1967. 1, 2

[46] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, pages 1034–1041. IEEE, 2009. 1, 2

[47] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456, 2003. 1, 2

[48] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009. 2

[49] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ICMI*, pages 517–524. ACM, 2013. 1, 2

[50] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524. ACM, 2013. 1, 2

[51] Jessica L Tracy, Richard W Robins, and Roberta A Schriber. Development of a facs-verified set of basic and self-conscious emotion expressions. *Emotion*, 9(4):554, 2009. 3

[52] http://mplab.ucsd.edu. The MPLab GENKI Database. 3

[53] Kai Wang, Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Cascade attention networks for group emotion recognition with face, body and image cues. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 640–645, New York, NY, USA, 2018. ACM. 3

[54] Kyoko Yamamoto and Naoto Suzuki. The effects of social interaction and personal relationships on facial expressions. *Journal of Nonverbal Behavior*, 30(4):167–179, 2006. 3

[55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 5

[56] Hengchin Yeh, Sean Curtis, Sachin Patil, Jur van den Berg, Dinesh Manocha, and Ming Lin. Composite agents. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 39–47. Eurographics Association, 2008. 4

[57] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL (Volume 1: Long Papers)*, pages 2236–2246, 2018. 2

[58] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019. 2, 3, 6, 7, 11, 12

[59] Jian Zheng, Sudha Krishnamurthy, Ruxin Chen, Min-Hung Chen, Zhenhao Ge, and Xiaohua Li. Image captioning with integrated bottom-up and multi-level residual top-down attention for game scene understanding. *arXiv preprint arXiv:1906.06632*, 2019. 4

[60] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012. 2

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 3