Emotion Recognition for Multiple Context Awareness

Dingkang Yang[1,2], Shuai Huang[1,2], Shunli Wang[1,2], Yang Liu[1], Peng Zhai[1,2], Liuzhen Su[1,2], Mingcheng Li[1,2], and Lihua Zhang[1,2,3,4(✉)]

[1] Academy for Engineering and Technology, Fudan University, Shanghai, China
lihuazhang@fudan.edu.cn
[2] Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China
[3] Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
[4] AI and Unmanned Systems Engineering Research Center of Jilin Province, Changchun, China

Abstract. Understanding emotion in context is a rising hotspot in the computer vision community. Existing methods lack reliable context semantics to mitigate uncertainty in expressing emotions and fail to model multiple context representations complementarily. To alleviate these issues, we present a context-aware emotion recognition framework that combines four complementary contexts. The first context is multimodal emotion recognition based on facial expression, facial landmarks, gesture and gait. Secondly, we adopt the channel and spatial attention modules to obtain the emotion semantics of the scene context. Inspired by sociology theory, we explore the emotion transmission between agents by constructing relationship graphs in the third context. Meanwhile, we propose a novel agent-object context, which aggregates emotion cues from the interactions between surrounding agents and objects in the scene to mitigate the ambiguity of prediction. Finally, we introduce an adaptive relevance fusion module for learning the shared representations among multiple contexts. Extensive experiments show that our approach outperforms the state-of-the-art methods on both EMOTIC and GroupWalk datasets. We also release a dataset annotated with diverse emotion labels, Human Emotion in Context (HECO). In practice, we compare with the existing methods on the HECO, and our approach obtains a higher classification average precision of 50.65% and a lower regression mean error rate of 0.7. The project is available at https://heco2022.github.io/.

Keywords: Emotion recognition · Context understanding

Di. Yang and S. Huang—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19836-6_9.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13697, pp. 144–162, 2022.
https://doi.org/10.1007/978-3-031-19836-6_9

1 Introduction

Understanding human emotion plays an essential role in daily life as emotion recognition has been applied in various complicated fields, such as medical care [10], human-computer interaction [12], and robotics [76]. Benefiting from the excellent performance of deep learning technologies in processing diverse signals [9,26,35,37,64,71], many researchers [6,38,50,55,58,59,70] have improved emotion recognition by combining diverse modalities (*e.g.*, face, audio, and language) from the recognized agent. Nevertheless, it is difficult to obtain the complete modalities from the different data domains, especially in practical applications where simplicity and practicality are the goals. In this paper, we analyse a wider view at the visual level to infer human emotion instead of focusing on the agent only.

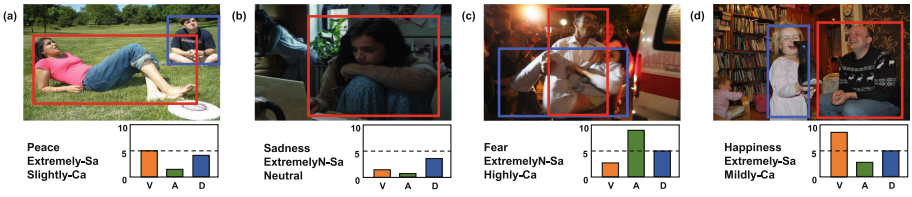

Fig. 1. Examples of agents in four contexts. The red rectangles present the recognized agent while the blue rectangles present the others. Discrete categories and continuous dimensions labels of emotion shown under the images. VAD means emotional state space: *Valence*, *Arousal*, and *Dominance*. (Color figure online)

Recently, emotion recognition that combines the agent's expression with the emotion semantics of context has received considerable attention [30,31,41,42,72]. Researches in context awareness inspire us to explore meaningful contexts from images and video frames to perceive emotion. There are some interesting examples. In Fig. 1(a) (*Explicit context*), the woman is lying on the grass with a flexible posture, whose emotion tends to be peaceful. Emotion sociology works [19,54,57] demonstrate that emotion is the maintenance and change of the relationship between agents and their scene. In Fig. 1(b) (*Scene context*), the performance of expression recognition in dark scene is limited and poor. However, the emotional state reflected in the surrounding environment is consistent with the agent, and it can be inferred from the scene context that the girl might be in negative emotion. Furthermore, the emotion transmission between multiple agents in the same scene can also affect the emotion change of the recognized agent. In Figure 1(c) (*Surrounding agent context*), the man rushes to the ambulance with an injured woman in his arms. Due to the woman's condition, the man feels fear. Moreover, inspired by emotion psychology studies [2,11,43], we consider emotion cues provided by implicit representation, such as agent-object interaction. In Fig. 1(d) (*Agent-object context*), the man feels happy when he sees his daughter have a good time with the hairdryer. The interaction between the girl and the hairdryer is beneficial for understanding the man's emotion. Cognitive scientists [17,40,47] state that humans exist in a society whose emotions can be affected by different

contexts directly or indirectly. Learning the multimodal representations from various contexts will effectively improve emotion recognition performance.

In summary, our primary contributions are the following: (1) We present a novel context-aware emotion recognition framework from a psychological and sociological perspective, which incorporates four context information. (2) We propose an adaptive relevance fusion module that focuses on the interactions among diverse contexts and adaptively assigns higher weights to beneficial contexts. (3) We release HECO, a new dataset for emotion recognition in context. The HECO is annotated with discrete and continuous emotion labels and promotes a more reasonable perception of human emotion.

2 Related Work

Uni/Multimodal Emotion Recognition. Isolated modalities, such as facial expression [74], voice [14], body gesture [44] and biological signal [3], have been concerned in prior emotion recognition works. Recently, multimodal emotion recognition [38,50,55,59] has been a hot issue, where researchers are incorporating multiple modalities to perform emotion analysis. Mainstream multimodal fusion strategies are classified as data-level [29,34], feature-level [52,70], and decision-level fusion [6,16]. In contrast, we propose a two-phase model-level fusion strategy with cross-context fusion and adaptive fusion. Our strategy reinforces the shared representations among multiple contexts in the interaction and assigns appropriate weights to the contexts based on their contributions.

Context-Aware Emotion Recognition. There have been several attempts at context-aware emotion recognition in recent years. Kosti *et al.* [30] propose the task of emotion understanding in context and build a two-stream Convolutional Neural Network (CNN) that combines the body and the semantic information from the scene. Zhang *et al.* [72] utilize the region proposal network to extract scene semantics as node features, and then construct an emotion graph through Graph Convolutional Network (GCN) to infer emotion. Lee *et al.* [31] use the attention mechanism to find relevant context cues in the scene after the hidden face. Mittal *et al.* [41,42] adopt a multiplicative fusion to combine information from various modalities and context interpretations. Hoang *et al.* [25] propose an extra reasoning stream to quantify the interaction between primary agents and objects. The aforementioned methods sub-optimally explore the emotion relationships between agents and the effect of agent-object interactions. In comparison, the four contexts proposed by our method are more complementary and synergistic in emotion recognition.

3 Proposed Method

3.1 Context 1: Explicit Multimodal Context

In the real world, the form of human emotion expression is usually multimodal. These modalities include facial expression [65,67], body posture [44,63], gesture [36,49], and walking style. It is helpful to infer emotion by integrating various

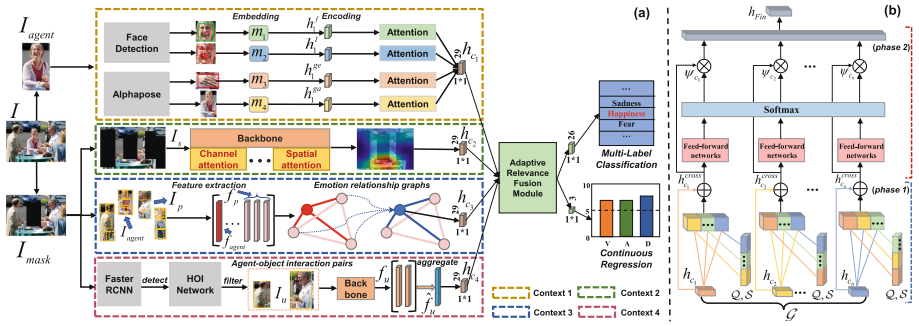

Fig. 2. (a) The proposed framework. We first extract features from face expression, facial landmarks, gesture and gait via respective neural networks to obtain $\boldsymbol{h}_1^f$, $\boldsymbol{h}_1^l$, $\boldsymbol{h}_1^{ge}$ and $\boldsymbol{h}_1^{ga}$. Then we fuse these features to obtain $\boldsymbol{h}_{c_1}$ in context 1. In contexts 2, 3 and 4, $\boldsymbol{h}_{c_2}$, $\boldsymbol{h}_{c_3}$ and $\boldsymbol{h}_{c_4}$ are obtained from different inputs via the corresponding context awareness models. Immediately, the Adaptive Relevance Fusion (ARF) module fuses all features and learns the multimodal representation $\boldsymbol{h}_{Fin}$. Finally, two separate branches perform the emotion classification and regression tasks. (b) The overall architecture of the ARF module contains two phases: cross-context fusion and adaptive fusion. (Color figure online)

modalities of emotion information [20,38,55]. To make full use of these emotion cues, as shown in the tawny-bordered branch of Fig. 2(a), context 1 utilizes diverse modalities as $\boldsymbol{m}_n$ from the recognized agent to extract multimodal representations such as the facial expression, facial landmarks, gesture and gait, which are defined as $\boldsymbol{H}_1 = \left\{\boldsymbol{h}_1^f, \boldsymbol{h}_1^l, \boldsymbol{h}_1^{ge}, \boldsymbol{h}_1^{ga}\right\}$. Formally, $\boldsymbol{m}_n$ from images or video frames I are encoded through respective neural network structure and the feature extraction process as follows:

$$\boldsymbol{h}_1^n = \mathcal{F}\left(\boldsymbol{m}_n; \boldsymbol{w}_n\right), \forall \boldsymbol{h}_1^n \in \boldsymbol{H}_1, \tag{1}$$

where $\boldsymbol{w}_n$ denotes the network parameters. Concretely, the ResNet-18 [22] is used to encode facial expression $\boldsymbol{m}_1$ to obtain the vector $\boldsymbol{h}_1^f \in \mathbb{R}^d$ from the fully connected layer. Concurrently, we employ three independent dense layers with a GeLU activation [24] to extract the features $\boldsymbol{h}_1^l, \boldsymbol{h}_1^{ge}$, and $\boldsymbol{h}_1^{ga}$ for facial landmarks $\boldsymbol{m}_2$, gesture $\boldsymbol{m}_3$, and gait $\boldsymbol{m}_4$, which have the identical dimension. Based on the different importance of these modalities, we propose a multimodal attention network to obtain the total vector $\boldsymbol{h}_{c_1} \in \mathbb{R}^d$ of context 1:

$$\mu_1^n = tanh(\boldsymbol{w}_{\mu n} \cdot \boldsymbol{h}_1^n + \boldsymbol{b}_{\mu n}), \tag{2}$$

$$\boldsymbol{h}_{c_1} = \sum_{n=1}^{N} \mu_1^n \odot \boldsymbol{h}_1^n, \tag{3}$$

where $\boldsymbol{w}_{\mu n} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_{\mu n} \in \mathbb{R}^{d \times 1}$ are the learnable parameters. The coefficient μ_1^n dynamically adjusts the contribution of each modality to the final representation of context 1.

3.2 Context 2: Scene Context

Exploring the surrounding semantics that affects the agent in a scene is indispensable for understanding human emotion [30,42,72]. For example, the input I includes semantic components composed of the *wine glass*, *dinner plate* and *sunny day* in Fig. 2(a). These factors may contain the emotion outpouring of the recognized agent. However, previous studies [31,41,42] have only masked the recognized agent's parts (*e.g.*, face or body), which can bring potential ambiguity of emotion generated by other agents in the same scene. To tackle this issue, our key idea is to mask all agents in input I to generate scene image I_s, which is expressed as:

$$I_{\mathrm{s}} = \begin{cases} I(i,j) & \text{if } I(i,j) \notin \mathrm{bbox}_{\mathrm{agent}} \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{4}$$

where $\mathrm{bbox}_{\mathrm{agent}}$ denotes the bounding box of the agent.

Inspired by visual attention [66], we utilize a scene-aware learning strategy based on the Channel Attention Module (CAM) and Spatial Attention Module (SAM) to capture the scene semantics that reflect emotion cues. The learning strategy is expected to make the model focused on the event context that effectively affects the agent's emotion. To encode the features in context 2, the ResNet-18 [22] is used to obtain the scene semantic vector $\boldsymbol{h}_{c_2} \in \mathbb{R}^d$ from the fully connected layer. The backbone is initialized by using the Places365-Standard [73], labelled with scene semantic categories. Concretely, we alternately insert the CAM and SAM in the eight residual blocks of the backbone. Given an intermediate feature map $\boldsymbol{x} \in \mathbb{R}^{c \times w \times h}$ as input, the CAM utilizes the global average pooling operation to infer a 1D channel attention map $M^c_{avg} \in \mathbb{R}^{c \times 1 \times 1}$, and the SAM utilizes global max pooling operation to infer a 2D spatial attention map $M^s_{max} \in \mathbb{R}^{1 \times w \times h}$. The overall attention process can be summarized as $\boldsymbol{x}_c = \left(\sigma\left(\delta M^c_{avg}(\boldsymbol{x})\right)\right) \otimes \boldsymbol{x}$ and $\boldsymbol{x}_s = \left(\sigma\left(\delta M^s_{max}(\boldsymbol{x})\right)\right)\boldsymbol{x}$, respectively, where $\otimes$ is channel-wise multiplication, $\delta(\cdot)$ is ReLU activation and $\sigma(\cdot)$ is sigmoid function. During multiplication, the channel attention values are broadcasted along the spatial dimension, and vice versa.

3.3 Context 3: Surrounding Agent Context

Motivated by emotion sociology studies [19,40,54,57], we find that surrounding agents with different intensities of emotion arousal and expression can help infer the primary agent's emotion. Nevertheless, previous works [23,42,69] mainly describe various interaction forces between agents as a single system. These methods are limited, which perfunctorily model the interaction distance and proximity between agents. Distinct from them, we aim to thoroughly explore the influence of surrounding agents' emotions on the recognized agent's expression.

Inspired by inductive learning [62], our core strategy is to construct dynamic graph structure to model the emotion relationships between agents. As shown in the blue-bordered branch of Fig. 2(a), we define the recognized agent's image as I_{agent} and the surrounding agents' images set as $\mathcal{I}_p = \left\{I^i_p\right\}, i = 1, ..., n$ by bounding boxes. After that, the conceptual node features $\boldsymbol{f}_{agent} \in \mathbb{R}^{d_s}$ and

$\boldsymbol{F}_p = \left\{ \boldsymbol{f}_p^i \in \mathbb{R}^{d_s} \right\}, i = 1, ..., n$ are extracted by final pooling layer in the ResNet-50 [22], respectively. Meanwhile, considering that individuals have different influences on emotion transmission [19], we assign different weights of emotion intensity to surrounding agents by performing attention. Formally, we calculate the emotion transfer coefficient e_i with LeakyReLU to measure the effect of each $\boldsymbol{f}_p^i$ on $\boldsymbol{f}_{agent}$, denoted as $e_i = \alpha([\boldsymbol{w}_a \boldsymbol{f}_{agent} \| \boldsymbol{w}_p \boldsymbol{f}_p^i])$, where $\|$ represents concatenation. The parameters $\boldsymbol{w}_a, \boldsymbol{w}_p \in \mathbb{R}^{d \times d_s}$ and the linear projection mapping $\alpha(\cdot)$ learn the emotion relationship. After performing the normalization via the softmax function, the final coefficient a_i is computed as $a_i = \frac{exp(e_i)}{\sum_{j \in \mathcal{N}} exp(e_j)}$, where $\mathcal{N}$ means the surrounding nodes. In practice, we set $K = 3$ to use the multi-head attention to realize the fusion of surrounding node features. The final weighted average feature $\boldsymbol{h}_{c_3} \in \mathbb{R}^d$ is obtained as follows:

$$\boldsymbol{h}_{c_3} = \sigma \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{i \in \mathcal{N}} a_i^k \boldsymbol{w}_p^k \boldsymbol{f}_p^i + \boldsymbol{w}_a^k \boldsymbol{f}_{agent} \right). \tag{5}$$

3.4 Context 4: Agent-Object Context

Emotion psychology researchers [4,11,43,45,53] emphasize that the interactions of agents' actions with objects induce emotion arousal of the primary agent in the scene. More colloquially, the context of interactions between surrounding agents and objects can trigger emotion cues that subliminally affect change in the emotion of the recognized agent, like the girl having fun with the hairdryer in Fig. 1(d), and the smiling man with the cup in Fig. 2(a). These interactions facilitate the outpouring of positive emotions by the recognized agents. Motivated by the above observations, our insight is to adopt an aggregation strategy to model the context of surrounding agent-object interactions and thus learn indirect representations.

More concretely, as shown in the red-bordered branch of Fig. 2(a), drawing on the success of Human-Object Interaction (HOI) task (detect the interactions between a human and object pair, then localize them) [18], we first define the bounding box set of agent-object interaction pairs obtained by input I_{mask} via the HOI Network [60] as $\mathcal{I}_u = \left\{ I_u^j \right\}, j = 1, .., m$. Subsequently, the pre-trained ResNet-50 [22] on ImageNet [15] separately encodes the interaction regions $\mathcal{I}_u$ to obtain the intermediate features as $\boldsymbol{F}_u = \left\{ \boldsymbol{f}_u^j \in \mathbb{R}^{d_s} \right\}, j = 1, .., m$. Immediately, the proposed aggregation strategy models the emotion semantics for different interaction pairs via learning dynamic weights:

$$\beta_u^j = \boldsymbol{U}^T (\boldsymbol{w}_u^j \cdot \boldsymbol{f}_u^j + \boldsymbol{b}_u^j), \tag{6}$$

$$\gamma_u^j = \frac{exp(\beta_u^j)}{\sum_{k=1}^{m} exp(\beta_u^k)}, \tag{7}$$

$$\hat{\boldsymbol{f}}_u = \sum_{j=1}^{m} \gamma_u^j \odot \boldsymbol{f}_u^j, \tag{8}$$

where $\boldsymbol{U} \in \mathbb{R}^{d_s \times 1}$, $\boldsymbol{w}_u^j \in \mathbb{R}^{d_s \times d_s}$, and $\boldsymbol{b}_u^j \in \mathbb{R}^{d_s \times 1}$ are the learnable parameters. After that, we perform a projection transformation on $\hat{\boldsymbol{f}}_u$ to obtain $\boldsymbol{h}_{c_4} \in \mathbb{R}^d$.

3.5 Feature Fusion and Learning Strategies

Considering the complementarity of diverse contexts and different levels of contributions, we propose an Adaptive Relevance Fusion (ARF) module to learn effective shared representations of contexts. As shown in Fig. 2(b), the ARF module consists of two phases: cross-context fusion and adaptive fusion. The cross-context fusion phase (phase 1) focuses on the interactions among different contexts, potentially adapting streams from one context to another. Note that our fusion strategy can be extended to diverse contexts. In this paper, we take the feature adaptation process about learning from $\boldsymbol{h}_{c_2}$ (context 2) to $\boldsymbol{h}_{c_1}$ (context 1) as an example to describe the details. Inspired by [61], the ARF module first embeds $\boldsymbol{h}_{c_1}$ into a space denoted as $\mathcal{G}_{c_1} = LN\left(\boldsymbol{h}_{c_1}\right)\boldsymbol{W}_{\mathcal{G}_{c_1}}$, while embedding $\boldsymbol{h}_{c_2}$ into two spaces denoted as $\mathcal{Q}_{c_2} = LN\left(\boldsymbol{h}_{c_2}\right)\boldsymbol{W}_{\mathcal{Q}_{c_2}}$ and $\mathcal{S}_{c_2} = LN\left(\boldsymbol{h}_{c_2}\right)\boldsymbol{W}_{\mathcal{S}_{c_2}}$, respectively, where $\boldsymbol{W}_{\mathcal{G}_{c_1}}$, $\boldsymbol{W}_{\mathcal{Q}_{c_2}}$, $\boldsymbol{W}_{\mathcal{S}_{c_2}} \in \mathbb{R}^{d\times d}$ are embedding weights, and LN means layer normalization. Attention weights are obtained by applying the softmax function to dot product of $\mathcal{G}_{c_1}$ and $\mathcal{Q}_{c_2}$. The information dissemination from cross-context interaction is defined as:

$$\boldsymbol{Z}^{cross}_{c_2\to c_1} = softmax(\mathcal{G}_{c_1}\mathcal{Q}^T_{c_2})\mathcal{S}_{c_2} \in \mathbb{R}^{d\times d}. \tag{9}$$

Immediately, the forward computation is expressed as:

$$\boldsymbol{h}^{cross}_{c_2\to c_1} = LN\left(\boldsymbol{h}_{c_1}\right) + \boldsymbol{Z}^{cross}_{c_2\to c_1}. \tag{10}$$

Assuming that the total set of context features as $\boldsymbol{H}_c = \{\boldsymbol{h}_{c_i}\}, i = 1, ..., n$, then the final interactions received by target context 1 as $\boldsymbol{h}^{cross}_{c_1} = \prod_{i=2}^{n} \boldsymbol{h}^{cross}_{c_i\to c_1}$, where $\prod$ denotes concatenation operator as $[\cdot\|\cdot]$ between features.

The adaptive fusion phase (phase 2) provides optimal fusion weights for each context to highlight the potent contexts while suppressing the weaker ones. Formally, we learn the attention weights through the respective feed-forward networks with a GeLU activation [24] denoted as $\psi_{c_i} = \mathcal{F}(\boldsymbol{h}^{cross}_{c_i}; \boldsymbol{w}_{c_i}), i = 1, ..., n$, where $\boldsymbol{w}_{c_i}$ are the network parameters. The softmax function makes the sum of these attentions to be 1, *i.e.*,$\sum_i \psi_{c_i} = 1$. After that, we perform element-wise multiplication of the learnable attention and the corresponding input. All outputs are concatenated and then fed to linear projection parametrized by $\boldsymbol{w}_\theta$ to obtain the feature $\boldsymbol{h}_{Fin}$, which is defined as follows:

$$\boldsymbol{h}_{Fin} = \sigma\left(\boldsymbol{w}_\theta \cdot \prod_{i=1}^{n} \psi_{c_i} \odot \boldsymbol{h}^{cross}_{c_i}\right). \tag{11}$$

Finally, two separate branches follow the fully connected layers, one for the discrete classification task and the other for the continuous regression task.

We use the MultiLabel-SoftMarginLoss as the classification loss of discrete categories, which is expressed as L_{disc}. The loss function of the continuous dimensions regression is formulated as $L_{cont} = \frac{1}{C}\sum_{k\in C}\left(\hat{y}_k - y_k\right)^2$, where y_k is the ground-truth of the continuous dimension regression, $\hat{y}_k$ is the output of VAD [39] dimensions and C is the number of channel dimensions. Therefore, the total training loss is defined as: $L_{comb} = \lambda_{disc}L_{disc} + \lambda_{cont}L_{cont}$, where λ_{disc} and λ_{cont} are the trade-off coefficients.

4 Datasets

EMOTIC. EMOTIC [30] dataset contains 23,571 images of 34,320 annotated people in uncontrolled environments. These images are annotated for 26 discrete categories and 3 continuous dimensions of emotion, with multiple labels assigned to each image. The standard partition of the dataset is 7:1:2.

GroupWalk. GroupWalk [42] dataset consists of 45 videos that were captured using stationary cameras in 8 real-world settings. The annotations consist of the following discrete labels: *Angry*, *Happy*, *Neutral*, and *Sad*. The standard partition of the dataset is 8.5:1.5.

HECO. HECO dataset consists of images from the HOI [8,21] datasets, film clips, and images from the Internet. The dataset contains a total number of 9,385 images and 19,781 annotated agents. These image samples contain rich context information and diverse agent interaction behaviours. To improve the robustness of models trained on the HECO, we add about 2% fuzzy images and 5% images with occlusion for agents. The dataset is randomly split into training (70%), validation (10%), and testing (20%) sets. The annotation process involves 3 psychologists and 10 graduate students. The annotation is performed blindly and independently, and we utilize the majority voting rule to determine the final labels. The superiority of HECO is that it combines two types of emotion labels. For discrete categories, we annotate with eight categories, including *Surprise*, *Excitement*, *Happiness*, *Peace*, *Disgust*, *Anger*, *Fear*, and *Sadness*. For continuous dimensions, we use the emotional state model of VAD [39], and annotate the *Valence* (V), *Arousal* (A) and *Dominance* (D) of agents on a scale of 1–10. Inspired by emotion sociology studies [19,57], we also design the novel *Self-assurance* (Sa) and *Catharsis* (Ca) labels. These labels describe the degree to which the agents interact with each other and adapt to the context.

5 Implementation Details

5.1 Data Processing

To generate I_{agent} and I_p from I via the bounding boxes, we use the pedestrian tracking method RobustTP [7] for the GroupWalk and the annotation information in the EMOTIC and HECO, respectively. For I_{agent} in context 1, we utilize the face detector [75] to implement face detection and clipping. The facial bounding boxes are used to get the facial input $\boldsymbol{m}_1$ and resize it to 64×64. We extract a 136-dimensional vector $\boldsymbol{m}_2 \in \mathbb{R}^{136}$ obtained through facial landmarks. We adopt the Alphapose [68] to obtain 18 modified gesture coordinates and 26 gait coordinates. In this case, the coordinates of key points are used to compute the 1D gesture vector $\boldsymbol{m}_3 \in \mathbb{R}^{36}$ and the 1D gait vector $\boldsymbol{m}_4 \in \mathbb{R}^{52}$. Then, the raw I is masked via the bounding boxes of I_{agent} to produce I_{mask}, and via the bounding boxes of I_{agent} and I_p to produce I_s in context 2. In context 3, the crop operation of the same size is performed from the middle and four

Table 1. Discrete classification results on the EMOTIC dataset.

Category	Kosti *et al.* [30]	Zhang *et al.* [72]	Lee *et al.* [31]	Mittal *et al.* [41]	Ours		Category	Kosti *et al.* [30]	Zhang *et al.* [72]	Lee *et al.* [31]	Mittal *et al.* [41]	Ours	
					L_{disc}	L_{comb}						L_{disc}	L_{comb}
Peace	22.35	30.68	19.55	**35.72**	25.5	26.24	Affection	26.47	**47.52**	22.36	38.55	41.61	37.66
Esteem	17.86	12.05	15.38	**25.75**	21.98	20.29	Anticipation	57.31	63.2	52.85	60.73	62.75	**63.31**
Engagement	86.69	**87.31**	73.71	86.23	74.69	75.23	Confidence	**80.33**	74.83	72.68	68.12	72.22	74.42
Happiness	58.92	72.9	53.73	80.45	83.58	**85.25**	Pleasure	46.72	48.37	34.12	67.31	67.26	**67.68**
Excitement	78.05	72.68	70.42	80.75	85.64	**86.56**	Surprise	22.38	8.44	17.46	19.6	25.31	**27.03**
Sympathy	15.23	19.45	14.89	16.74	24.7	**25.87**	Doubt/ Confusion	31.88	19.67	26.07	**38.43**	23.44	24.96
Disconnection	20.64	23.17	22.01	28.73	27.64	**28.95**	Fatigue	8.87	12.93	6.29	19.35	32.35	**33.58**
Embarrassment	3.05	1.58	1.88	10.31	9.63	**10.57**	Yearning	9.22	9.86	4.84	**15.08**	10.88	11.12
Disapproval	16.14	12.64	15.37	18.55	23.41	**23.52**	Aversion	7.44	6.81	3.26	11.33	13.19	**15.28**
Annoyance	15.26	12.33	14.42	24.68	28.98	**29.02**	Anger	11.24	11.27	12.88	14.69	15.47	**17.84**
Sensitivity	9.05	4.74	6.94	13.94	22.53	**24.89**	Sadness	18.69	23.9	17.75	40.26	46.75	**47.8**
Disquietment	19.57	17.66	10.84	**22.14**	19.36	21.17	Fear	15.7	6.15	7.47	16.99	36.06	**36.68**
Pain	9.46	8.22	8.16	14.68	18.26	**19.27**	Suffering	17.67	23.71	14.85	**48.05**	45.37	46.74
							mAP	27.93	28.16	23.85	35.28	36.87	**37.73**

Table 2. Discrete classification results on the GroupWalk dataset.

Category	Kosti *et al.* [30]	Zhang *et al.* [72]	Lee *et al.* [31]	Mittal *et al.* [41]	Ours
Anger	57.65	51.92	45.18	68.85	**70.54**
Happy	71.32	63.37	56.59	72.31	**72.38**
Neutral	43.1	40.26	39.32	50.34	**52.54**
Sad	61.24	58.15	52.96	70.8	**71.42**
mAP	58.33	53.43	48.51	65.58	**66.72**

Table 3. Discrete classification results on the HECO dataset.

Category	Kosti *et al.* [30]	Zhang *et al.* [72]	Lee *et al.* [31]	Mittal *et al.* [41]	Ours	
					L_{disc}	L_{comb}
Surprise	28.45	34.87	24.27	**38.37**	38.04	38.12
Excitement	42.16	45.74	37.97	48.59	53.2	**55.04**
Happiness	62.82	63.26	55.81	66.53	67.26	**69.16**
Peace	51.64	54.17	47.57	55.97	57.23	**57.31**
Disgust	45.37	49.43	41.74	50.48	52.28	**54.95**
Anger	40.76	45.22	38.39	51.29	53.04	**53.43**
Fear	32.74	35.67	30.51	**40.81**	40.08	40.27
Sadness	22.53	27.28	20.92	32.65	34.17	**36.94**
mAP	40.81	44.46	37.15	48.09	49.41	**50.65**

corners to obtain sub-images I_{crop} from I_{agent} and I_p, which consider the semantic shape information. In context 4, we use the Faster RCNN [51] to generate available bounding boxes of agents and objects, which are then fed into the HOI network [60] to obtain the agent-object pairs with interaction relations. After that, we obtain the union bounding boxes that contain the specific agent-object pairs based on the coordinate transformation. The above images are resized to 224×224. All backbones employed are empirical.

5.2 Training Details and Evaluation Metric

Our method is built on the Pytorch toolbox [48], and all models are trained on four Nvidia Tesla V100 GPUs. We extend the proposed method to video by averaging the prediction vectors of all frames in the GroupWalk. For the EMOTIC, GroupWalk, and HECO, the training batch sizes and epochs are $\{64, 1, 32\}$ and $\{120, 75, 100\}$, respectively. The Adam optimizer [27] is adopted for network optimization with an initial learning rate of $\{1e^{-3}, 1e^{-4}, 2e^{-3}\}$. To train the previous models on the HECO, we use the implementation provided by [30]. Limited by open source, we re-implemented [31,41,72] based on the reported details. Following [30,72], we use the Average Precision (AP) to evaluate classification results. To evaluate regression results, we employ the Error Rate (ER).

6 Experimental Results

6.1 Comparison with State-of-the-Art Methods

Discrete Classification Results. In Tables 1, 2 and 3, we report the AP scores for all categories and mean AP (mAP). Our method achieves the best results of 37.73%, 66.72% and 50.65% on the EMOTIC, GroupWalk and HECO, respectively, significantly improving 2–8% over the prior methods. Concretely, we observe that the AP scores of some categories is generally low, such as *Sensitivity* from the EMOTIC and *Sadness* from the HECO. However, our method remains competitive in these categories. Furthermore, we train different models by the combined loss L_{comb} and the discrete loss L_{disc} respectively for testing. Except for the categories *Affection* and *Esteem* on the EMOTIC, the results of L_{comb} are superior, showing that combining different emotion expressions is beneficial in depicting emotions.

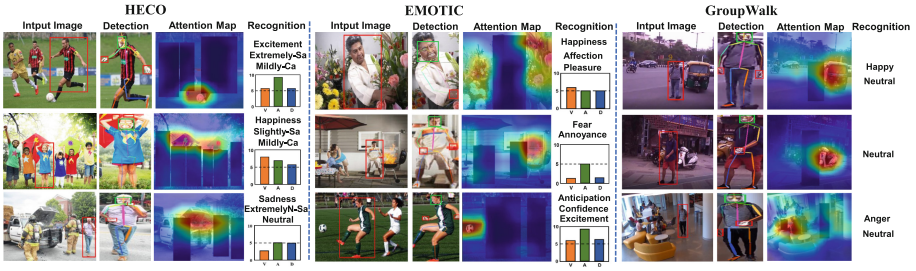

Fig. 3. Visualization results. We respectively show three examples of classification and regression results from the HECO, EMOTIC, and GroupWalk datasets. Column 1 is the input images marked with the recognized agent. Column 2 shows the facial expression, facial landmarks, gesture and gait extracted from the agent. Column 3 shows the corresponding attention maps. Column 4 shows the recognized emotion labels.

Table 4. Continuous regression results on the EMOTIC and HECO datasets.

Method	Dataset	Valence	Arousal	Dominance	mER	Dataset	Valence	Arousal	Dominance	mER
Kosti *et al.* [30] (L_{cont})	EMOTIC	1.0	1.5	0.8	1.1	HECO	0.9	1.3	0.8	1.0
Kosti *et al.* [30] (L_{comb})		0.9	1.2	0.9	1.0		0.9	1.2	**0.6**	0.9
Zhang *et al.* [72] (L_{cont})		0.8	1.6	1.2	1.2		0.9	1.1	1.0	1.0
Zhang *et al.* [72] (L_{comb})		0.7	1.0	1.0	0.9		**0.6**	1.1	0.7	0.8
Ours (L_{cont})		**0.6**	1.3	0.8	0.9		0.8	1.0	**0.6**	0.8
Ours (L_{comb})		0.8	**0.9**	**0.7**	**0.8**		0.7	**0.8**	**0.6**	**0.7**

Continuous Regression Results. Table 4 shows the evaluation results for the continuous dimensions using the mean ER (mER). We compare methods for supporting regression task on the EMOTIC and HECO. Note that the GroupWalk has only discrete emotion labels. On both datasets, our method outperforms the previous methods [30,72] with the lowest mER. We notice that the mER of L_{comb} is lower than that of L_{cont} for each method, which indicates that learning discrete classification contributes to infer emotional state of continuous dimensions. Additionally, all methods have lower mER on the HECO than EMOTIC, which mainly benefit from diverse sample sources and rich agent interaction instances in the HECO to assist the models in recognizing emotion.

6.2 Visualization and Analysis

Case Study of Multimodal Attention. To verify the effectiveness of the proposed multimodal attention in context 1, we visualize several detected samples in Fig. 4(a). Each recognized agent in the sample has the multimodal representations, including facial expression, facial landmarks, gesture and gait. We calculate the average L2-normalization of the vector attention for multiple modalities. The heat map matrix represents the attention intensity of each modality from different samples. For instance, in the first row, the girl's clear face conveys joy more clearly than the other modalities, so the facial expression and landmarks have the higher weights. In contrast, the lady's face in the third row is incomplete, but we can reasonably infer emotion by her body language (the gesture about caresses). As a result, the gesture feature obtains the highest weight. The above observations show that the multimodal attention can effectively learn the dynamic contribution of different modalities to the final representation.

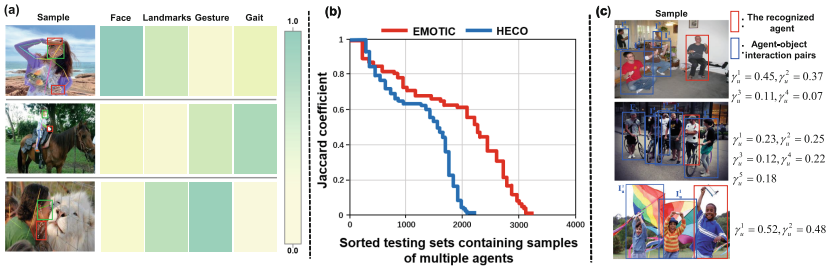

Fig. 4. (a) Heat map of attention weights for multiple modalities from different samples in context 1. The darker colour indicates the higher weights. (b) Jaccard coefficients for samples containing multiple agents in the testing set (sorted). The higher values denote the existence of emotion co-occurrence between the recognized agent and surrounding agents in context 3. (c) For context 4, we provide dynamic weight analysis of agent-object interaction pairs in the aggregation strategy. γ_u^j come from Eq. 7.

Emotion Semantic Capture of Scene. Figure 3 presents the visualization results of three examples for each dataset. The attention maps show the scenes' emotion semantics learned from the network. For example, in the samples of the second row (*middle*) and the last row (*left*), the semantic context of *fire* and *crashed car* are interpreted, implying *fear* and *sadness*, respectively. Meanwhile, although agents in the first row (*middle*) and the second row (*left*) have same discrete emotion *Happiness*, different continuous emotions inspire us better understand agents' adaptability in the scene and physiological arousal level. Moreover, other contexts can contribute to making predictions when some agents' features are difficult to extract due to occlusion or ambiguity, such as gesture.

Emotion Co-occurrence. To explain the emotion relationship between agents, we calculate the Jaccard coefficient [46] for each sample of multiple agents containing annotations on the EMOTIC and HECO datasets, respectively. Concretely, we define the set of predicted categories (the recognized agent) in each sample as S_{pred} and the set of the ground truth of surrounding agents in same sample as $\left(S_a^1, S_a^2, ..., S_a^n\right) \subseteq S_a$. n denotes the number of agents. The Jaccard coefficient is computed as $(S_a \cap S_{pred})/(S_a \cup S_{pred})$. In Fig. 4(b), we observe that over 64% of the samples in both datasets have values above 0.6, *i.e.*, the emotional states of the recognized agents are consistent with or similar to the surrounding agents. This observation proves that learning the emotion relationship between agents in the same scene can assist in inferring the primary emotion.

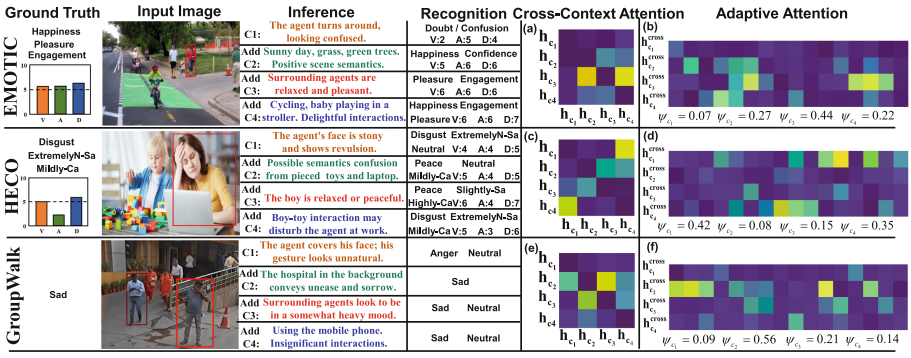

Fig. 5. Complementarity analysis. (*Left*) We select a sample from each of three datasets and then perform the recognition by gradually adding context network branches. (*Right)* When the above samples use four context branches, we plot the attention matrix (a), (c) and (e) for the cross-context fusion phase and the corresponding attention matrix (b), (d) and (f) for the adaptive fusion phase.

Analysis of Aggregation Strategy. In Fig. 4(c), we show several samples to understand the effect of different agent-object interactions. γ_u^j are the weight coefficients obtained by the corresponding interaction pairs I_u^j in the sample through the aggregation strategy. Some interesting observations are as follows. The effect of the agent-object interaction pairs is significant when the recognized agent and surrounding agents are involved in the same or similar event, and vice versa. For example, in the first row, I_u^1 and I_u^2 in the game playing case correspond to higher coefficients $\gamma_u^1 = 0.45$ and $\gamma_u^2 = 0.37$. In contrast, I_u^4 in the food-taking case has the lowest coefficient. A similar pattern is found in the second row, where the coefficient for $\gamma_u^3 = 0.12$ in the case of drinking water is the lowest. Furthermore, we find that the effect tendency of agent-object interaction pairs to generally conform to a gradual decay outwards centred on the recognized agent. These observations align with the psychology theories [11,53].

Complementarity Analysis. To prove the complementary recognition ability of four contexts (referred to as C_1, C_2, C_3, and C_4), we select a sample from each of three datasets and then perform inference by gradually adding context network branches. Figure 5 shows the dynamic recognition results during the addition of context branches. On the EMOTIC, positive semantics of scene from C_2, relaxed other agents from C_3, and pleasant interactions from C_4 gradually remove the emotion ambiguity recognized by C_1 only. On the HECO, the semantics of the intrusive boy-toy interaction from C_4 enhances the emotion judgment of *Disgust* in C_1. That is, the superior performance benefits from combining multiple context information to complement the emotion cues.

Table 5. Ablation study results on the EMOTIC, GroupWalk, and HECO datasets.

Model design	Dataset						Model design	Dataset					
	EMOTIC		GroupWalk		HECO			EMOTIC		GroupWalk		HECO	
	mAP	mER	mAP	mER	mAP	mER		mAP	mER	mAP	mER	mAP	mER
Full (ours)	**37.73**	**0.8**	**66.72**	–	**50.65**	**0.7**	C_3 (GCNs) [28]	35.25	1.2	63.94	–	48.84	1.1
C_1	22.51	1.6	44.76	–	37.27	1.4	C_3 (Depth) [32]	36.39	1.0	65.19	–	49.02	0.9
$C_1 + C_2$	29.23	1.1	54.42	–	41.93	1.0	Concatenation [52]	30.47	1.2	59.87	–	43.6	1.1
$C_1 + C_3$	27.56	1.3	57.36	–	39.62	1.1	Multiplication [41]	36.52	0.9	65.24	–	50.22	0.7
$C_1 + C_4$	26.29	1.2	52.09	–	37.93	1.2	ARF (phase 1)	36.27	0.9	65.33	–	49.73	0.8
$C_1 + C_2 + C_3$	36.18	0.8	64.34	–	48.07	0.8	ARF (phase 2)	34.65	1.0	64.13	–	47.51	0.9
$C_1 + C_2 + C_4$	34.93	0.9	60.27	–	47.25	0.8	C_1 (OpenFace [1]+OpenPose [5])	37.45	0.8	66.39	–	50.28	0.8
$C_1 + C_3 + C_4$	33.45	1.0	62.61	–	44.23	0.9	C_4 (R-FCN [13]+HOI-Net [33])	37.52	0.9	66.45	–	50.34	0.7
C_2 (Mask Face)	35.57	1.0	64.34	–	47.71	1.0	VGG19 [56] (C_1, C_2)+Res101 [22] (C_3, C_4)	37.76	0.9	66.68	–	50.87	0.7
C_2 (Mask Body)	37.02	0.9	65.12	–	49.46	0.8	Res34 [22] (C_1, C_2)+Res152 [22] (C_3, C_4)	36.83	0.8	65.85	–	50.24	0.8

Interpreting Cross-Context Fusion Attention. We plot the attention matrices for the cross-context fusion phase when using four contexts for the above samples illustrated in Fig. 5(a, c, e). The attention matrix shows the adaptation and interaction of features on the vertical axis to features on the horizontal axis. The brighter areas represent the higher correlation among context features that can collectively improve emotion expression. For a reasonable example, on the HECO, the correct recognition bring about by the C_1 and C_4 branches correspond to the brighter areas in Fig. 5(c) *w.r.t.* $\boldsymbol{h}_{c_1 \to c_4}^{cross}$ and $\boldsymbol{h}_{c_4 \to c_1}^{cross}$.

Interpreting Adaptive Fusion Attention. In Fig. 5(b, d, f), we visualize the attention weights of the adaptive fusion phase for the above samples when using four contexts. The vertical axis of the attention matrix denotes the features in different contexts, and the horizontal axis denotes the partial dimension. A higher number of bright areas in a feature indicate that the feature contributes more to emotion recognition. As an example on the GroupWalk, since the branch of C_2 successfully captures the *sadness* shown by the agent walking out of the hospital, which corresponds to the brightest areas contained by $\boldsymbol{h}_{c_2}^{cross}$ in Fig. 5(f). Obviously, $\psi_{c_2} = 0.56$ is the highest weight coefficient.

6.3 Ablation Study

We perform thorough ablation study of all components to demonstrate the effectiveness and robustness of our method. Table 5 shows the results for discrete categories and continuous dimensions on three datasets.

Effectiveness of Context Branches. For context branches, we keep context 1, which only captures information from the recognized agent itself, and then gradually remove other context networks. Note that the best result is to combine four context branches (Full). For discrete categories, contexts 1 and 2 are the most competitive of two combinations. Such advantage comes from the fact that most of the samples on the EMOTIC and HECO have rich scene elements. The equally good results of contexts 1 and 3 on the GroupWalk may benefit from vast agent flows, which is consistent with the observations of [41,42]. Furthermore, similar results are observed for continuous dimensions.

Different Masking Strategies. When using full branches, we provide two strategies to replace masking all agents of I_s in context 2, *i.e.*, masking only the face of the agent proposed in [31] and only the body of the agent proposed in [41]. The results show that the previous strategies of masking only the part of the recognized agent clearly hurt the model's performance, which suggests that masking all agents is essential.

Modelling of Interaction Between Agents. For context 3, we present the alternatives to evaluate the rationality of the chosen structure. More concretely, we replace the branch in context 3 with the GCNs-based [28] and Depth-based [32] methods used in [41] to model the social dynamics of interactions between agents. Our method outperforms the alternative versions as the emotion relationships of surrounding agents are modelled through attention weights explicitly, rather than simply as a system.

Different Fusion Strategies. To show the advantages of the ARF module, we perform comparison experiments with the concatenation [52] and multiplicative fusion [41] of the final features from different contexts. The results show that the ARF module is more competitive, proving that capturing correlations across contexts can provide effective multimodal representations. Moreover, we retain one phase of the ARF module separately for testing. It is observed that the fusion mechanisms from both phases provide the indispensable contributions.

Effect of Detectors. We replace the detectors with alternative components (OpenFace [1] and OpenPose [5] in C_1, R-FCN [13] and HOI-Net [33] in C_4) to explore whether there is an effect on the model's performance. The results in Table 5 show that replacing detectors has a slight effect on the mER and that the errors in the mAP scores are both less than 0.35. The above observations prove that our method is robust, *i.e.*, the detectors barely affect performance.

Analysis of Backbone CNNs. In addition, we use different backbones to implement the proposed framework. Table 5 shows that a deeper network structure does not necessarily obtain better results, *i.e.*, the performance improvement does not depend entirely on the backbones.

7 Conclusion

In this paper, we propose a novel context-aware emotion recognition framework, which employs four meaningful context branches to understand human emotion in a boosting and synergistic manner. Inspired by emotion sociology and psychology, we explore emotion-rich representations from contexts at the visual level to advance the development of effective visual-only driven emotion recognition applications. Moreover, learning multimodal shared representations through the proposed adaptive relevance fusion module allows for extending our approach to more contexts. Numerous qualitative and quantitative analyses clearly demonstrate the superiority of our approach.

Acknowledgements. This work is supported by National Key R&D Program of China (2021ZD0113502, 2021ZD0113503), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103) and National Natural Science Foundation of China under Grant (82090052).

References

1. Baltrušaitis, T., Robinson, P., Morency, L.P.: OpenFace: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
2. Barrett, L.F., Mesquita, B., Gendron, M.: Context in emotion perception. Curr. Dir. Psychol. Sci. **20**(5), 286–290 (2011)
3. Bos, D.O., et al.: EEG-based emotion recognition. The influence of visual and auditory stimuli, vol. 56, no. 3, pp. 1–17 (2006)
4. Calhoun, C., Solomon, R.C.: What is an emotion?: classic readings in philosophical psychology (1984)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
6. Castellano, G., Kessous, L., Caridakis, G.: Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868, pp. 92–103. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85099-1_8

7. Chandra, R., Bhattacharya, U., Roncal, C., Bera, A., Manocha, D.: RobustTP: end-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In: ACM Computer Science in Cars Symposium, pp. 1–9 (2019)
8. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 381–389. IEEE Computer Society (2018)
9. Chen, Z., Li, B., Xu, J., Wu, S., Ding, S., Zhang, W.: Towards practical certifiable patch defense with vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15148–15158 (2022)
10. Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T.: Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun. **50**(6), 487–503 (2008)
11. Cornelius, R.R.: The Science of Emotion: Research and Tradition in the Psychology of Emotions. Prentice-Hall, Inc., Upper Saddle River (1996)
12. Cowie, R., et al.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
13. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems 29 (2016)
14. Davidson, R.J., Sherer, K.R., Goldsmith, H.H.: Handbook of Affective Sciences. Oxford University Press, Oxford (2009)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
16. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Acted facial expressions in the wild database. Australia, Technical report TR-CS-11 2, 1, Australian National University, Canberra (2011)
17. Frijda, N.H.: Emotion, cognitive structure, and action tendency. Cogn. Emot. **1**(2), 115–143 (1987)
18. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8359–8367 (2018)
19. Gordon, S.L.: The sociology of sentiments and emotion. In: Social psychology, pp. 562–592. Routledge (2017)
20. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. J. Netw. Comput. Appl. **30**(4), 1334–1345 (2007)
21. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
23. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
24. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)
25. Hoang, M.H., Kim, S.H., Yang, H.J., Lee, G.S.: Context-aware emotion recognition based on visual relationship detection. IEEE Access **9**, 90465–90474 (2021)
26. Huang, H., et al.: CMUA-watermark: a cross-model universal adversarial watermark for combating deepfakes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 989–997 (2022). https://doi.org/10.1609/aaai.v36i1.19982. https://ojs.aaai.org/index.php/AAAI/article/view/19982

27. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2014)
28. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
29. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: data level fusion strategy for hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2103–2111 (2018)
30. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE Trans. Pattern Anal. Mach. Intell. **42**(11), 2755–2766 (2019)
31. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10143–10152 (2019)
32. Li, Z., Snavely, N.: MegaDepth: learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050 (2018)
33. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: parallel point detection and matching for real-time human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 482–490 (2020)
34. Liu, K., Gebraeel, N.Z., Shi, J.: A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. IEEE Trans. Autom. Sci. Eng. **10**(3), 652–664 (2013)
35. Liu, S., et al.: Efficient universal shuffle attack for visual object tracking. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2739–2743 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747773
36. Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., Zhao, G.: iMiGUE: an identity-free video dataset for micro-gesture understanding and emotion analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10631–10642 (2021)
37. Liu, Y., Liu, J., Zhao, M., Li, S., Song, L.: Collaborative normality learning framework for weakly supervised video anomaly detection. IEEE Trans. Circuits Syst. II Express Briefs **69**(5), 2508–2512 (2022). https://doi.org/10.1109/TCSII.2022.3161061
38. Lu, Y., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: IJCAI, vol. 15, pp. 1170–1176. Citeseer (2015)
39. Mehrabian, A.: Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies, vol. 2. Oelgeschlager, Gunn & Hain, Cambridge (1980)
40. Mesquita, B., Boiger, M.: Emotions in context: a sociodynamic model of emotions. Emot. Rev. **6**(4), 298–302 (2014)
41. Mittal, T., Bera, A., Manocha, D.: Multimodal and context-aware motion perception model with multiplicative fusion. IEEE MultiMedia **28**, 67–75 (2021)
42. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: Context-aware multimodal emotion recognition using Frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14234–14243 (2020)
43. Musch, J., Klauer, K.C.: The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Psychology Press, Brighton (2003)

44. Navarretta, C.: Individuality in communicative bodily behaviours. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) Cognitive Behavioural Systems. LNCS, vol. 7403, pp. 417–423. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34584-5_37
45. Niedenthal, P.M., Ric, F.: Psychology of Emotion. Psychology Press, Brighton (2017)
46. Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of Jaccard coefficient for keywords similarity. In: Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1, pp. 380–384 (2013)
47. Ochsner, K.N., Gross, J.J.: The cognitive control of emotion. Trends Cogn. Sci. **9**(5), 242–249 (2005)
48. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
49. Piana, S., Stagliano, A., Odone, F., Verri, A., Camurri, A.: Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047 (2014)
50. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 873–883 (2017)
51. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28, pp. 91–99 (2015)
52. Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A.N., Prasad, R.: Emotion recognition using acoustic and lexical features. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
53. Ruckmick, C.A.: The psychology of feeling and emotion (1936)
54. Schachter, S., Singer, J.: Cognitive, social, and physiological determinants of emotional state. Psychol. Rev. **69**(5), 379 (1962)
55. Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., Bartlett, M.: Multiple kernel learning for emotion recognition in the wild. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 517–524 (2013)
56. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
57. Stets, J.E.: Current emotion research in sociology: advances in the discipline. Emot. Rev. **4**(3), 326–334 (2012)
58. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference Meeting on Association for Computational Linguistics, vol. 2019, p. 6558. NIH Public Access (2019)
59. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176 (2018)
60. Ulutan, O., Iftekhar, A., Manjunath, B.S.: VSGNet: spatial attention network for detecting human object interactions using graph convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13617–13626 (2020)
61. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
62. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
63. Wallbott, H.G.: Bodily expression of emotion. European J. Soc. Psychol. **28**(6), 879–896 (1998)

64. Wang, S., Yang, D., Zhai, P., Chen, C., Zhang, L.: TSA-NET: tube self-attention network for action quality assessment. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4902–4910 (2021)
65. Wang, W., et al.: Comp-GAN: compositional generative adversarial network in synthesizing and recognizing facial expression. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 211–219 (2019)
66. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
67. Xie, S., Hu, H., Wu, Y.: Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recogn. **92**, 177–191 (2019)
68. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose Flow: efficient online pose tracking. In: BMVC (2018)
69. Yeh, H., Curtis, S., Patil, S., van den Berg, J., Manocha, D., Lin, M.: Composite agents. In: Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 39–47 (2008)
70. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016)
71. Zhai, P., Luo, J., Dong, Z., Zhang, L., Wang, S., Yang, D.: Robust adversarial reinforcement learning with dissipation inequation constraint (2022)
72. Zhang, M., Liang, Y., Ma, H.: Context-aware affective graph reasoning for emotion recognition. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 151–156. IEEE (2019)
73. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Analy. Mach. Intell. **40**(6), 1452–1464 (2017)
74. Zhu, J., Luo, B., Zhao, S., Ying, S., Zhao, X., Gao, Y.: IExpressNet: facial expression recognition with incremental classes. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2899–2908 (2020)
75. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)
76. Ziemke, T.: On the role of emotion in biological and robotic autonomy. BioSystems **91**(2), 401–408 (2008)