Neurocomputing 620 (2025) 129195

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# TACL: A Trusted Action-enhanced Curriculum Learning Approach to Multimodal Affective Computing

Tan Yu [a], Jingjing Wang [a,b,*], Jiamin Luo [a], Jiawen Wang [a], Guodong Zhou [a]

[a] School of Computer Science and Technology, Soochow University, Suzhou, China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Suzhou, China

ARTICLE INFO

Communicated by M. Xu

*Keywords:*
Dynamic and temporal action information
Trusted learning
Curriculum learning
Multimodal Affective Computing

ABSTRACT

Previous studies on Multimodal Affective Computing (MAC) predominantly focus on leveraging language, acoustic, and facial information to identify human's affective states, which largely ignore the dynamic and temporal action information, despite such information being crucial for precisely inferring the affective states. In this way, this paper first attempts to consider the action information for MAC and further argues that exploiting the action information faces two key challenges, i.e., credibility and sparsity challenges. To this end, this paper proposes a new **T**rusted **A**ction-enhanced **C**urriculum **L**earning (TACL) approach to incorporate the action information for boosting MAC. Specifically, this approach designs two main components, i.e., the Trusted Curriculum Learning block and the Action-enhanced Vision Regulator, to address the above credibility challenge and sparsity challenge. Furthermore, a high-quality action-enhanced video dataset is constructed to evaluate TACL and detailed evaluations show the great advantage of TACL over the state-of-the-art baselines. Particularly, an interesting finding is observed that action information is more conducive to facilitating the recognition of negative emotions, which aligns with the intuition that humans prefer using actions more when expressing negative emotions.

## 1. Introduction

Multimodal Affective Computing (MAC), a fundamental task in the field of multimodal learning [1], seeks to harness various information encompassing spoken words (i.e., language), tones (i.e., acoustic), and facial attributes (i.e., visual) to achieve precise prediction of emotion or sentiment categories. Distinct from unimodal affective computing, MAC holds the distinct advantage of facilitating a holistic comprehension and discernment of human affective states. Thus, MAC has wide applications, including human–computer interaction [2], social media [3], and mental health [4].

The MAC task aims to predict the emotion or sentiment categories by leveraging multimodal signals, e.g., language, acoustic, and face. Previous studies can be divided into two paradigms: Fusion Paradigm [5–13] and Representation Paradigm [14–20]. For the fusion paradigm, Chen et al. [7] used an RNN-based model to perform modality fusion at the word level. Tsai et al. [6] introduce the Multimodal Transformer to address the data no-alignment and long-range dependencies issues in MAC, and Han et al. [8] proposed the Multi-Modal InfoMax which maximizes the Mutual Information in unimodal input pairs to promote multimodal fusion. For the representation paradigm, Pham et al. [16] learned robust joint representations by translating between modalities. Yu et al. [17] designed a label generation module to acquire independent unimodal supervisions and then joint-trained the multimodal and uni-modal tasks to learn the modal representation, and Hu et al. [21] propose a multimodal sentiment knowledge-sharing framework, which leverage features, labels and models to fully exploit the complementary knowledge behind the sentiment and emotion. Besides, a few studies [22–25] have realized the importance of action information and leverage such information for emotion detection, but limited to single-modal scenarios.

Different from all the above studies, this paper first attempts to consider the dynamic and temporal action information for the MAC task and proposes two inherent credibility and sparsity challenges in capturing such action information for MAC. As shown in Fig. 1, an interesting scenario wherein a boy ostensibly expresses acceptance towards veggies using the phrase "*That's good.*". Yet, a nuanced micro-action *shrug* subtly reveals his true emotion *disgust*. While the action information constitutes merely one fact of emotional judgment, it imparts invaluable emotional clues [26]. In light of these observations, this paper considers incorporating the action information for boosting

* Corresponding author.
*E-mail addresses:* tyu417@stu.suda.edu.cn (T. Yu), djingwang@suda.edu.cn (J. Wang), jmluo97@outlook.com (J. Luo), 20235227102@stu.suda.edu.cn (J. Wang), gdzhou@suda.edu.cn (G. Zhou).

https://doi.org/10.1016/j.neucom.2024.129195
Received 30 April 2024; Received in revised form 17 November 2024; Accepted 14 December 2024
Available online 24 December 2024

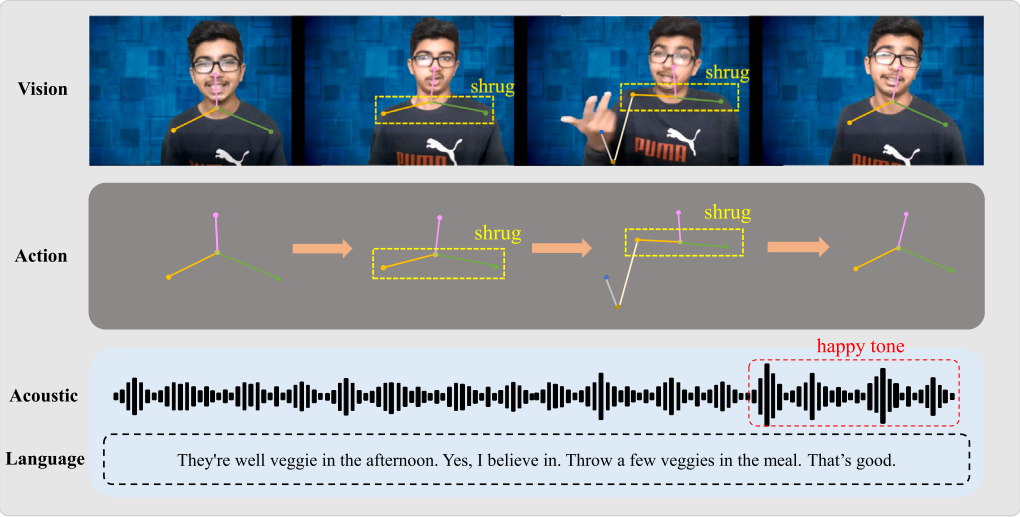



**Fig. 1.** A multimodal example with the *disgust* label from our constructed action-enhanced video dataset to illustrate the importance of action information in MAC, where the "*shrug*" action (yellow boxes) indicates the *disgust* emotion.

MAC. Especially, we believe that capturing such information faces the following two main challenges.

For one thing, this paper argues that action information has higher credibility in comparison with other modalities (e.g., language, acoustic, and face) in terms of expressing affective states (namely the "credibility challenge" for short). Actions generally offer an inherently direct way for the communication of emotions, thereby precisely reflecting true affective states, as reported by [27]. The empirical evidence depicted in Fig. 1 exemplifies this phenomenon, where predictions based on language (e.g., "*That's good.*"), acoustic (e.g., *happy tones*) and facial (e.g., *smile face*) information render a positive emotion for the boy, whereas the action reveals a clear indication of negative emotion (e.g., *disgust*) through the micro-action (i.e., *shrug*) information. As actions are unconscious human behaviors, challenging to dissimulate or control, we believe that it is rather credible in expressing truly affective states. Therefore, a well-behaved approach should highlight the high credibility characteristic of action information for the MAC task.

For another, this paper argues that action information is sparse, making it difficult to precisely capture the affective states expressed by actions (namely the "sparsity challenge" for short). Intuitively, individuals are always lazy to adopt action to express affective states, leading to the sparsity of the action information, but once the action is used, even modest actions possess the capacity to effectively convey corresponding affective states. Still taking Fig. 1 as an example, the micro-action *shrug* exhibits subtle characteristics and short duration but betrays the clear positive emotion, making it rather difficult yet important to recognize its affective states. Therefore, a better-behaved approach should also seek to mitigate the sparsity issue of action information for MAC.

In this paper, we introduce the curriculum learning mechanism [28] and propose a **T**rusted **A**ction-enhanced **C**urriculum **L**earning (TACL) approach to tackle the aforementioned challenges. Particularly, this TACL approach first integrates the trusted learning [29] into curriculum learning for designing a Trusted Curriculum Learning (TCL) block to address the credibility challenges of the action information for MAC. Furthermore, an Action-enhanced Vision Regulator (AVR) is designed to further mitigate the sparsity issue. Overall, the main contributions of our work are summarized as follows:

- We present a pioneering effort to consider the dynamic and temporal action information for MAC, and propose two credibility and sparsity challenges in capturing such action information.
- We propose a new TACL approach to address the above two credibility and sparsity challenges for MAC. This approach is the first attempt to integrate both curriculum learning and trusted learning techniques.
- We construct an action-enhanced video dataset to evaluate our TACL approach and detailed evaluations demonstrate the superior performance of TACL over the state-of-the-art baselines, justifying the importance of the action information and the effectiveness of our approach in capturing such information.

## 2. Related work

In this paper, we leverage curriculum learning to design our approach. In the following, we will present recent research advancements in curriculum learning and discuss how our approach differs from previous methods. Inspired by the human learning process, curriculum learning trains models by gradually adjusting the difficulty and complexity of training data, which assists models in learning more effectively. Bengio et al. [28] are the first to introduce the concept of curriculum learning and explore why curriculum learning aids model learning, and Graves et al. [30] use curriculum learning to help the neural networks optimize effectively.

Recently, some studies [31,32] attempt to employ curriculum learning to compute the difficulty degrees of different multimodal samples and achieve promising results on various multimodal tasks, which is inspirational to our TACL approach.

In summary, unlike prior studies, this paper first attempts to integrate the trusted learning [29] into curriculum learning and design a difficulty scoring function (see Eq. (4)) to address the credibility challenge of action for boosting MAC.

## 3. Trusted action-enhanced curriculum learning

In this paper, the MAC task is formulated as follows. Given the multimodal signal $\mathbf{I} = \{\mathbf{I}^l, \mathbf{I}^s, \mathbf{I}^f, \mathbf{I}^a\}$, where $\mathbf{I}^l, \mathbf{I}^s, \mathbf{I}^f, \mathbf{I}^a$ denote the unimodal raw sequence from the language, acoustic, facial and action modalities, respectively. The goal of the MAC task is to predict the emotion category and the sentiment polarity of the multimodal signal **I**. In this paper, we propose a **T**rusted **A**ction-enhanced **C**urriculum **L**earning (TACL) approach to capture and utilize action information effectively, thereby boosting the performance of MAC. The overall framework of TACL is shown in Fig. 2, which mainly comprises the Trusted Curriculum Learning (TCL) Block and Action-enhanced MAC Block.

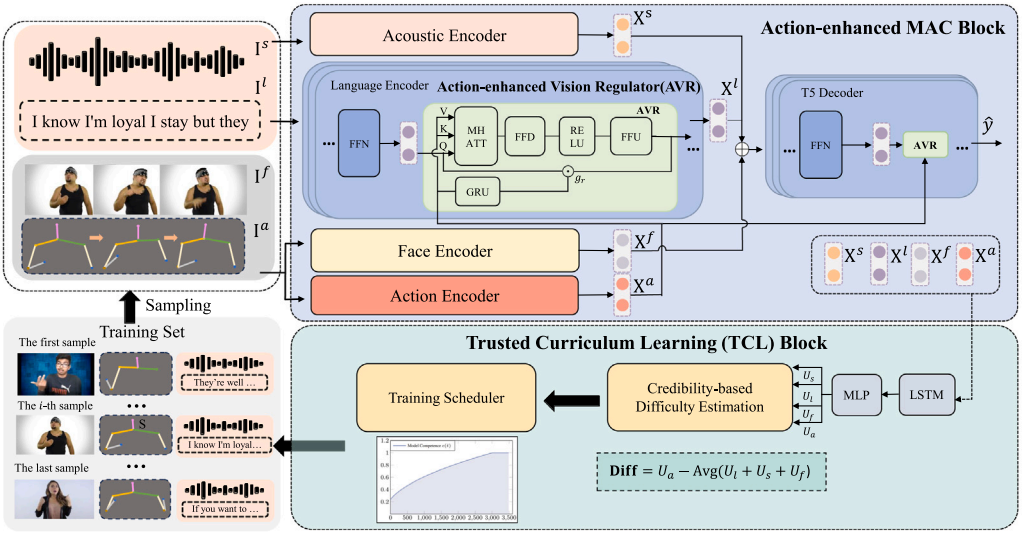



**Fig. 2.** The overall architecture of our proposed TACL approach to the MAC task.

### 3.1. Action-enhanced MAC block

The Action-enhanced MAC block is designed to make predictions from the multimodal signals. Specifically, we first use Librosa [33], T5 Encoder, OpenFace [34] and HRNet [35] as Acoustic Encoder, Language Encoder, Face Encoder, and Action Encoder to obtain the acoustic representation $\mathbf{X}^s$, language representation $\mathbf{X}^l$, face representation $\mathbf{X}^f$, and action representation $\mathbf{X}^a$, respectively.

Then, the language representation $\mathbf{X}^l$ is concatenated with the acoustic representation $\mathbf{X}^s$ and face representation $\mathbf{X}^f$ to input into the T5 Decoder to get the predictions. In addition, we elaborately design an Action-enhanced Vision Regulator (AVR) to incorporate the action information into the model.

**Action-enhanced Vision Regulator** is designed to capture the action information effectively and mitigate the sparsity challenge, which follows the feed-forward layer in each of several Transformer layers of T5. Specifically, to better leverage the role of action information, we introduce a retained gate, which decides how much proportion of the modality representation that does not fuse action information to be retained. The retained gate $g_r^i$ for the $i$th AVR is calculated as:

$$\begin{aligned} \mathbf{h}_a^i &= \mathrm{sGRU}(\mathbf{X}^a; \theta^{gru}) \\ g_r^i &= \sigma(\mathbf{W}_r^i \mathbf{h}_a^i) \end{aligned} \tag{1}$$

where $\mathbf{X}^a$ denotes the action representation. $\mathbf{h}_a^i$ is the end-state hidden vector obtained by encoding $\mathbf{X}^a$ using a single directional Gate Recurrent Unit(sGRU). $\sigma$ represents the sigmoid function. $\theta^{gru}$ are the parameters of the sGRU and $\mathbf{W}_r^i$ is the projection matrix.

After getting retained gate $g_r^i$, the output $\mathbf{F}_{out}^i$ of the $i$th AVR is denoted as follows:

$$\begin{aligned} \mathbf{F}_d^i &= \mathrm{MHATT}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ \mathbf{F}_h^i &= \mathrm{FFU}(\mathrm{RELU}(\mathrm{FFD}(\mathbf{F}_d^i))) \\ \mathbf{F}_{out}^i &= g_r^i \odot \mathbf{F}_f + \mathbf{F}_h^i \end{aligned} \tag{2}$$

where $\mathbf{Q} = \mathbf{W}_q^i \mathbf{F}_f$, $\mathbf{K} = \mathbf{W}_k^i \mathbf{X}^a$ and $\mathbf{V} = \mathbf{W}_v^i \mathbf{X}^a$. $\mathbf{F}_f$ is the output vector of the Feed-Forward Network(FFN) in the transformer layer. $\mathbf{X}^a$ is the action representation. $\mathbf{W}_q^i, \mathbf{W}_k^i, \mathbf{W}_v^i$ are the projection matrices. MHATT represents multi-head attention. FFU and FFD refer to the Feed-Forward up-project and Feed-Forward down-project layers, respectively. RELU is the relu activation function, and $\odot$ means component-wise multiplication.

### 3.2. Trusted curriculum learning block

In this paper, the Trusted Curriculum Learning (TCL) Block is designed to fully leverage the high credibility of action information for boosting the performance of MAC.

Specifically, this paper integrates trusted learning into curriculum learning and proposes a TCL block, which consists of two main components: credibility-based difficulty estimation and training scheduler for highlighting the higher credibility of action information.

**Credibility-based Difficulty Estimation** aims to calculate the uncertainty for each modality in each sample (i.e., Credibility Calculation) and designs a difficulty scoring function (i.e., Difficulty Scoring Function) for curriculum learning. The uncertainty for a modality in a sample is defined as the uncertainty of classifying the sample using only that modal information (calculated by Eq. (3)). In our approach, the uncertainty and credibility are inversely proportional (i.e., the lower the uncertainty, the higher the credibility).

- **Credibility Calculation.** Inspired by [36], we employ the Dempster–Shafer evidence theory [37] and Subjective Logic [38] to calculate modality credibility. Specifically, for a $K$ classification task (e.g., our emotion detection task), as shown in Fig. 2 TCL block, we input a certain modality representation (e.g., language representation $\mathbf{X}^l$) into LSTM followed by a multilayer perceptron (MLP). Then, RELU is employed as the activation function of the output layer to get the output probability values, and these probability values are regarded as the evidence $e = [e_1, \ldots, e_K]$ as proposed by Sensoy et al. [36]. Subsequently, we adopt Subjective Logic [38] to assign a belief mass $b_k$ to each class $k = 1, \ldots, K$ and provide an overall modality uncertainty mass $U$ based on the evidence. Here, these $K + 1$ mass values are all non-negative and their sum is 1, i.e., $U + \sum_{k=1}^{K} b_k = 1$, where $U \geq 0$ and $b_k = \frac{e_k}{\sum_{i=1}^{K} e_i + 1}$. On this basis, the modality uncertainty $U$ (e.g., uncertainty of language modality $U_l$) is derived as follows:

$$U = \frac{K}{\sum_{i=1}^{K} e_i + 1} \tag{3}$$

From Eq. (3), we can observe that when category number $K$ is fixed, the smaller sum of evidence for all categories means the greater the modality uncertainty (i.e., the lower credibility). During optimizing our TACL approach, we leverage the Dirichlet distribution to compute the distribution of the evidence $e = [e_1, \ldots, e_K]$ for computing modality uncertainty $U$. More specifically, the Subjective Logic is used to connect the evidence $e = [e_1, \ldots, e_K]$ to the parameters of the Dirichlet distribution $\alpha = [\alpha_1, \ldots, \alpha_K]$, i.e., $\alpha_k = e_k + 1$, where the Dirichlet

**Table 1**
Statistics of our constructed Action-Enhanced Video (AEV) dataset.

| Data Setting | Emotion detection | | | | | Sentiment classification | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | #Anger | #Disgust | #Happy | #Neutral | #Sad | #Positive | #Neutral | #Negative | |
| Training | 348 | 136 | 632 | 891 | 355 | 632 | 891 | 839 | 2362 |
| Validation | 49 | 19 | 90 | 127 | 50 | 90 | 127 | 118 | 335 |
| Test | 102 | 38 | 180 | 255 | 102 | 180 | 255 | 242 | 677 |

distribution can be considered as the conjugate prior of the evidence $e = [e_1, \ldots, e_K]$ [39]. That is, a subjective opinion can be derived easily from the corresponding Dirichlet distribution using $b_k = (\alpha_k - 1)/H$, where $H = \sum_{i=1}^{K} e_i + 1 = \sum_{i=1}^{K} \alpha_i$ is the Dirichlet strength.

The detailed computation process for each modality uncertainty $U$ is illustrated in Training Stage 1 of algorithm 1.

• **Difficulty Scoring Function** is designed to determine the difficulty of each sample for performing curriculum learning.

Since the credibility of action modality is expected to be higher than that of language, acoustic, and facial modalities in this study, we calculate the difference **Diff** according to Eq. (4) where the larger **Diff** means this sample is more difficult. Here, the larger **Diff** indicates the larger $U_a$ (i.e., the lower credibility of action) compared to the other three modalities (i.e., language, acoustic, and face). This situation suggests that the classification results of this sample are easily misled by language, acoustic, and face, making it more difficult for the model to learn.

$$\textbf{Diff} = U_a - \text{Avg}(U_l + U_s + U_f) \tag{4}$$

where Avg represents the operation of taking the average. The difference **Diff** denotes the difficulty level of a sample. $U_a$, $U_l$, $U_s$, and $U_f$ denote the uncertainty of action, language, acoustic, and facial modalities of this sample, which is calculated by Eq. (3).

**Training Scheduler.** After sorting the training data using the difficulty scoring function in Eq. (4), the training scheduler is designed to determine when we should present harder data for training, and how much more data should be included. In this paper, we take advantage of the competence-based training scheduler proposed by [40] to train our TACL approach. Specifically, the model competence $c(t) \in (0, 1]$ at training step $t$ is defined using the following functional forms:

$$c(t) = \text{Min}(1, \sqrt[2]{t\frac{1 - c(0)^2}{T} + c(0)^2}) \tag{5}$$

where $c(0)$ and $T$ are the initial competence and the total number of training steps, respectively. Min is the minimum function.

As shown in Training stage 2 of Algorithm 1, at training time step $t$, a batch is sampled from the top $c(t)$ portions of the sorted training dataset to train the model. During training, the value of $c(t)$ increases from 0 to 1 (shown in Fig. 2), which means that our TACL approach gradually learns samples from simple to more challenging ones.

### 3.3. Action-enhanced optimization

To capture the action information effectively, our TACL approach is trained in two stages, as illustrated in Algorithm 1. Specifically, in the first stage, we employ the uncertainty loss $\mathcal{L}^{(u)}$ (proposed by Sensoy et al. [36]) as the optimization objective to calculate the uncertainty of action, language, acoustic, and face modalities $U_a$, $U_l$, $U_s$, and $U_f$. The uncertainty loss $\mathcal{L}^{(u)}$ is defined as:

$$\mathcal{L}^{(u)} = \sum_{i=1}^{N}(\sum_{j=1}^{K} y_{ij}(\psi(H_i) - \psi(\alpha_{ij})) + \lambda_t \mathbb{KL}[\mathbb{D}(p_i|\tilde{\boldsymbol{\alpha}}_i) \parallel \mathbb{D}(p_i|1)]) \tag{6}$$

The first term of $\mathcal{L}^{(u)}$ is designed to ensure that the correct label of each sample generates more evidence than other classes, and the second term of $\mathcal{L}^{(u)}$ is a regularization term to guarantee that less evidence generated for incorrect labels. $\psi(\cdot)$ is the digamma function. $y_{ij}$ denotes the expected class label of the $j$th class for the $i$th sample. $H_i$ is the Dirichlet strength of the $i$th sample, and $\alpha_{ij}$ is the parameter of the Dirichlet distribution of the $i$th sample for class $j$. $\mathbb{D}(\cdot)$ represents the Dirichlet distribution. And $\mathbb{KL}$ denotes the Kullback–Leibler divergence [41]. $p_i$ is the class assignment probabilities on a simplex, and $\lambda_t > 0$ is the balance factor, which gradually increases as the training progresses to allow the model to explore the parameter space and avoid premature convergence to the uniform distribution for misclassified samples. $\tilde{\boldsymbol{\alpha}}_i = y_i + (1 - y_i) \odot \boldsymbol{\alpha}_i$ is the adjusted parameter of the Dirichlet distribution to prevent the evidence of the ground truth class from being penalized to zero. $y_i$ is a one-hot vector encoding the ground-truth class $j$ of the $i$th sample with $y_{ij} = 1$ and $y_{ik} = 0$ for all $k \neq j$. $\boldsymbol{\alpha}_i =< \alpha_{i1}, \ldots, \alpha_{iK} >$ is the parameters of a Dirichlet distribution for the classification of a sample $i$.

In the second stage, we use the cross-entropy loss $\mathcal{L}^{(ce)}$ as the optimization objective for MAC to train the model.

$$\mathcal{L}^{(ce)} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log(\hat{y_{ij}}) \tag{7}$$

## 4. Experimental settings

### 4.1. Dataset construction and setting

In the literature, almost all existing public MAC datasets (e.g., MOSI [42], MOSEI [43], SIMS [44]) lack body action information. Therefore, in order to evaluate the effectiveness of our TACL to MAC, we construct an Action-Enhanced Video (AEV) dataset, which is composed of 3374 relatively long affective videos with an average length of 30 s. Specifically, this AEV dataset is constructed based on the released OMG-Emotion dataset [45] which contains body actions instead of merely human faces. Then, through manual selection, we remove the videos without actions from the original dataset for better evaluating the contributions of actions. Consequently, for language, we employ the Google Speech Recognition API[1] to get language from the videos. For acoustic and face, Librosa [33] and OpenFace [34] are used to extract acoustic and facial features, respectively. For action features, HRNet [35] is used to predict human body keypoints, and then the variations information of these keypoints between video frames is utilized to obtain dynamic and temporal action features by following Jhuang et al. [46]. After the dataset construction, we obtain the AEV dataset. On this basis, we build two classical tasks (i.e., emotion detection and sentiment classification) and randomly split the dataset into training, validation, and test sets in a ratio of 7:1:2. Detailed statistics w.r.t. two tasks in the AEV dataset are shown in Table 1. Here, for sentiment classification, we consider *happy* as a *positive* sentiment, *anger*, *sad*, and *disgust* as *negative* sentiments, and *neutral* as a *neutral* sentiment.

### 4.2. Baselines

We choose several advanced baselines in MAC to compare performance with our approach, described as follows. **SPC** [47] alleviates the problem of self-attention complexity and memory footprint. **MulT** [6] uses the directional pairwise crossmodal attention to attend to interactions between multimodal sequences. **MAG-BERT** [48] allows BERT to accept acoustic and visual data during fine-tune by applying Multimodal Adaptation Gate. **MISA** [49] learns effective modality

[1] https://cloud.google.com/speech-to-text/

**Algorithm 1** Action-Enhanced Optimization

**Input**: training dataset $\mathbf{D}^{train} = \{\mathbf{I}^l, \mathbf{I}^s, \mathbf{I}^f, \mathbf{I}^a\}$
**Output**: optimal model $M^*$

**Training Stage 1:** Computing the uncertainty $U_l$, $U_s$, $U_f$ and $U_a$ for each training sample in $\mathbf{D}^{train}$
**for** each sample $\mathbf{S}_i = \{\mathbf{I}_i^l, \mathbf{I}_i^s, \mathbf{I}_i^f, \mathbf{I}_i^a\}_{i=1}^N$ **do**
  Input $\mathbf{I}_i^l, \mathbf{I}_i^s, \mathbf{I}_i^f, \mathbf{I}_i^a$ to four encoders to get the language representation $\mathbf{X}_i^l$, acoustic representation $\mathbf{X}_i^s$, face representation $\mathbf{X}_i^f$, and representation $\mathbf{X}_i^a$
  Input $\mathbf{X}_i^l$, $\mathbf{X}_i^s$, $\mathbf{X}_i^f$, and $\mathbf{X}_i^a$ separately to four independent LSTM and MLP and train them with the uncertainty loss $\mathcal{L}^{(u)}$ in Eq. (6)
**end for**
Using the four trained LSTM and MLP to get the uncertainty $U_l$, $U_s$, $U_f$ and $U_a$ of the language, acoustic, facial, and action modalities by Eq. (3) for each sample $\mathbf{S}_i$
Sort each sample in $\mathbf{D}^{train}$ based on Eq. (4) to acquire $\hat{\mathbf{D}}^{train}$
**Training Stage 2:** Training model with TCL
Initialize the model competence $c(0)$ by Eq. (5), training step $t = 0$
**while** $t \neq T$ **do**
  Compute model competence $c(t)$ by Eq. (5)
  Uniformly sample a data batch $B(t)$ from the top $c(t)$ portions of $\hat{\mathbf{D}}^{train}$
  Input the data batch $B(t)$ into the model and optimize with $\mathcal{L}^{(ce)}$ in Eq. (7) for training the model $M$
  $t \leftarrow t + 1$
**end while**

**Table 2**
Results on our AEV dataset. w/o Action denotes removing the dynamic and temporal action information from TACL. w/o Language, w/o Acoustic, and w/o Face denote removing language, acoustic, and face information, respectively. w/o TCL and w/o AVR denote removing TCL block and AVR in TACL, respectively. w/o TL denotes removing trusted learning (TL) in TCL block.

| Approach | Emotion Detection | | | | | | | Sentiment Classification | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anger | Disgust | Happy | Neutral | Sad | **W-F1** | **Acc** | Positive | Neutral | Negative | **W-F1** | **Acc** |
| SPC | 18.99 | 20.82 | 38.26 | 52.76 | 25.82 | 37.97 | 39.38 | 37.58 | 44.46 | 43.25 | 42.20 | 42.34 |
| + Action | 3.79(↑) | 4.15(↓) | 3.95(↑) | 2.01(↓) | 0.82(↓) | **0.50**(↑) | **0.32**(↑) | 2.67(↑) | 2.37(↓) | 2.45(↑) | **0.69**(↑) | **0.50**(↑) |
| MulT | 25.06 | 9.67 | 42.75 | 49.70 | 28.76 | 38.74 | 39.97 | 41.65 | 43.16 | 43.55 | 42.90 | 42.84 |
| + Action | 6.36(↓) | 1.16(↓) | 1.49(↑) | 2.59(↑) | 4.08(↑) | **0.96**(↑) | **0.93**(↑) | 7.56(↓) | 0.46(↑) | 7.24(↑) | **0.75**(↑) | **1.34**(↑) |
| MAG-BERT | 27.47 | 20.07 | 21.75 | 51.18 | 34.28 | 35.49 | 38.04 | 31.53 | 39.96 | 49.69 | 41.20 | 42.54 |
| + Action | 10.69(↓) | 2.28(↑) | 14.52(↑) | 2.40(↑) | 16.09(↓) | **0.86**(↑) | **0.63**(↑) | 1.66(↓) | 3.16(↑) | 0.06(↓) | **0.72**(↑) | **0.89**(↑) |
| MISA | 18.37 | 16.13 | 37.66 | 52.35 | 11.34 | 35.11 | 37.01 | 33.35 | 44.44 | 45.52 | 41.88 | 42.16 |
| + Action | 4.58(↓) | 7.40(↑) | 3.10(↑) | 6.80(↓) | 19.33(↑) | **0.90**(↑) | **0.80**(↑) | 8.13(↑) | 3.18(↓) | 0.72(↓) | **0.70**(↑) | **0.68**(↑) |
| MMIM | 24.45 | 11.54 | 39.26 | 51.43 | 32.58 | 39.05 | 39.94 | 32.44 | 48.87 | 48.54 | 44.38 | 45.27 |
| + Action | 1.75(↓) | 11.54(↑) | 1.75(↑) | 0.48(↓) | 1.06(↓) | **0.51**(↑) | **0.83**(↑) | 5.37(↑) | 0.24(↓) | 0.07(↑) | **1.37**(↑) | **0.68**(↑) |
| UniMSE | 21.07 | 22.12 | 35.55 | 49.77 | 31.04 | 37.29 | 38.26 | 39.32 | 50.92 | 43.10 | 45.04 | 45.28 |
| + Action | 0.75(↓) | 7.57(↓) | 1.51(↓) | 0.81(↑) | 6.04(↑) | **0.28**(↑) | **0.88**(↑) | 12.20(↑) | 2.13(↓) | 4.96(↓) | **0.67**(↑) | **0.56**(↑) |
| **TACL(ours)** | 23.17 | 27.07 | 42.98 | 51.80 | 34.66 | **41.17** | **42.80** | 41.95 | 49.83 | 48.85 | **47.38** | **47.62** |
| w/o Action | 16.46 | 10.28 | 44.85 | 51.49 | 27.57 | 38.53 | 39.05 | 46.12 | 49.88 | 35.92 | 43.89 | 44.55 |
| w/o Language | 18.35 | 20.11 | 41.88 | 46.25 | 25.62 | 36.31 | 36.87 | 43.21 | 46.36 | 34.73 | 41.37 | 41.72 |
| w/o Acoustic | 20.22 | 21.55 | 42.34 | 46.91 | 28.83 | 37.53 | 37.91 | 37.08 | 47.54 | 40.22 | 42.14 | 43.09 |
| w/o Face | 24.43 | 25.10 | 40.02 | 48.77 | 25.13 | 37.89 | 38.42 | 40.66 | 46.90 | 41.33 | 43.24 | 43.77 |
| w/o TCL | 21.68 | 24.96 | 38.09 | 49.20 | 36.72 | 38.86 | 39.59 | 41.72 | 48.61 | 46.37 | 45.98 | 46.11 |
| w/o TL | 23.81 | 25.19 | 39.94 | 45.75 | 37.89 | 38.56 | 39.11 | 45.69 | 47.87 | 42.33 | 45.31 | 46.02 |
| w/o AVR | 26.80 | 24.74 | 37.06 | 47.23 | 31.28 | 37.78 | 38.82 | 40.71 | 42.90 | 46.72 | 43.68 | 43.77 |

representations by projecting each modality to modality-invariant and modality-specific space. **MMIM** [8] maintains task-related information through multimodal fusion by hierarchically maximizing the Mutual Information. **UniMSE** [21] conducts modality fusion at the syntactic and semantic levels and introduces contrastive learning for better capturing the difference and consistency between sentiments and emotions. This model is the state-of-the-art model in MAC task.

### 4.3. Implementation details and metrics

In our experiments, we re-implement all the baselines on the AEV dataset according to their open-source codes. For all baselines, we employ the same multimodal fusion strategy as the original models to incorporate the action information into the models. The hyper-parameters of these baselines are tuned according to the validation set for best performance. For a fair comparison, we conduct five independent runs and report the average performance of TACL and all the baselines. Due to the presence of class imbalance in the dataset, with a scarcity of samples in the *disgust* class, we performed data augmentation by tripling the number of *disgust* class samples in the training set. Following Hu et al. [21], we also use the pre-trained T5-Base[2] as the backbone of TACL. The hyper-parameters of our approach are also tuned according to the validation set. Specifically, for TACL, we set the batch size, $c(0)$, and $T$ to be 32, 0.24 and 3000, respectively. We use Adam optimizer as the optimizer and use the initial learning rate of 6e-5 for T5 fine-tuning, 1e-4 for AVR, and 1e-3 for other parameters. Moreover, t-test [50] is used to evaluate the significance of the performance difference between the two approaches. In addition, we use weighted average f1-score (W-F1) and accuracy (Acc) as the

[2] https://github.com/huggingface/transformers/tree/main/src/transformers/models/t5

evaluation metric. We also report F1 score per class. The F1, ACC and W-F1 are denoted as:

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{W-F1} &= \sum_{i=1}^{n} \frac{N_i}{N} \cdot \text{F1}_i \\
\text{ACC} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}
\end{aligned} \tag{8}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. $n$ is the total number of categories. $N_i$ is the number of samples in the $i$th category. $N$ is the total number of all samples. , and $\text{F1}_i$ is the F1 score of the $i$th category.

## 5. Results and discussions

### 5.1. Experimental results and analysis

Table 2 illustrates the comparative results for MAC on the AEV dataset. For all the baselines, we report their performance with and without using action information. From the table, we can observe that: **(1)** These baselines with additional action information can consistently achieve better performance than only using three modalities (i.e., language, acoustic, and face), which justifies the importance of action information for MAC. **(2)** However, after integrating action information, the performance of all baselines only shows a slight improvement for MAC. This is mainly due to that action information is sparse and challenging to capture, which encourages us to address the sparsity challenge of action information. **(3)** Our TACL approach outperforms all the baselines and achieves state-of-the-art performance. More importantly, compared to the state-of-the-art baseline UniMSE, our TACL approach improves W-F1 and Acc of emotion detection, W-F1 and Acc of sentiment classification by 3.60%, 3.66%, 1.67%, and 1.78%, respectively ($p$-value < 0.01). These results justify the appropriateness and effectiveness of our approach and encourage us to consider the action information in MAC.

Furthermore, we conduct a series of ablation studies to explore the contribution of each component in TACL, also as shown in Table 2. **(1) w/o Action.** To justify the effectiveness of the action information for MAC and the effectiveness of our approach in utilizing action information, we eliminate the action modality in TACL. Then, we observe that the performance drops more sharply ($p$-value < 0.01) compared to all the baselines. This indicates the effectiveness of action information for the MAC task and also demonstrates that our TACL approach can better mitigate the sparsity challenge of action information compared to other baselines. Interestingly, we observe that the performance for negative emotions, such as *anger*, *disgust*, and *sad*, decreases more significantly than the positive ones after removing the action information. This indicates that action information is more conducive to facilitating the recognition of negative emotions, which aligns with the intuition that humans prefer using actions more when expressing negative emotions. **(2) w/o Language, w/o Acoustic, and w/o Face.** w/o Language, w/o Acoustic, and w/o Face all exhibit inferior performance compared to TACL, which justifies that language, acoustic, and face modality information can all enhance the performance of the MAC task. **(3) w/o TCL.** To justify the effectiveness of the TCL block, we remove the TCL block, which means using random sampling from the training set to input the samples, and the results show a significant decline ($p$-value < 0.01) in performance. This demonstrates the effectiveness of TCL block and encourages us to integrate trusted learning and curriculum learning to address the credibility challenge of action information for improving the performance of MAC. What is more, the performance without TCL still outperforms UniMSE. This indicates that compared to the multimodal fusion layer in UniMSE, our AVR based on the retained gate and attention mechanism can better capture dynamic and temporal action information and mitigate sparsity challenges, resulting in improved performance. **(4) w/o AVR.** To justify the effectiveness of AVR, we remove AVR in TACL. We can observe that the performance decreases significantly ($p$-value < 0.01). This demonstrates that the AVR can mitigate the sparsity issue of action information. **(5) w/o TL.** To justify the effectiveness of trusted learning (TL), we remove TL in the TCL block, which means replacing Eq. (4) with a random difficulty scoring function for curriculum learning. Then, the results show a significant decrease ($p$-value < 0.01). This indicates the effectiveness of TL and encourages us to integrate TL into curriculum learning to address the credibility challenge for boosting the performance of MAC.

### 5.2. Convergence analysis

In order to validate that our model has been effectively optimized and achieved convergence, we save the training loss and model performance at each step. As depicted in Fig. 3, the loss value of TACL decreases until convergence, and the model performance of TACL continuously improves until convergence, which indicates that the model has been adequately trained. Besides, we also compared the convergence speed of our TACL model with UniMSE and TACL without action information. From Fig. 3, we can see that: **(1) TACL** demonstrates faster convergence compared to UniMSE, which justifies that our TACL approach can better utilize action information, leading to faster convergence. **(2) TACL** demonstrates faster convergence compared to w/o Action, which justifies that effectiveness of action information.

### 5.3. Is action really more credible and sparse?

In this paper, we argue that action information has higher credibility and is sparse for MAC. To justify our argument, we compare the single-modal uncertainty (Fig. 4(a)) and single-modal performance (Fig. 4(b)).[3]

Besides, in order to verify the robustness of our uncertainty computation, we randomly select training samples with various sizes to calculate uncertainty.

As shown in Fig. 4(a), across various training data settings, the uncertainty of the action modality remains consistently lower compared to the other modalities. This verifies that action information is more credible compared to other modalities. Particularly, we can observe a specific credibility ranking, i.e., action > face > acoustic > language. From Fig. 4(b), we can observe a significant performance gap for action modality compared to the other three modalities. This indicates that action information is sparse and difficult to capture, which encourages us to leverage AVR to mitigate the sparsity challenge.

Besides, it should be noted that the action credibility can be understood from the perspective of precision. For example, action is generally sparse (i.e., most actions do not contain affective elements), but once an action includes affective, it is generally real and could be seen as highly credible. As for the reason why other modalities, e.g., language, perform better than action, this is mainly because that language has stronger representation tools like T5 and LLMs than other modalities. Especially in the era of Large Language Models (LLMs), some recent studies [51] have also realized the modality credibility problem and pointed out that model predictions tend to be more biased towards language, ignoring other modalities. This also inspires us to consider how to address modality credibility bias in multimodal LLMs.

[3] For a fair comparison, we use the same model architecture (i.e., an LSTM followed by an MLP) for four modalities.

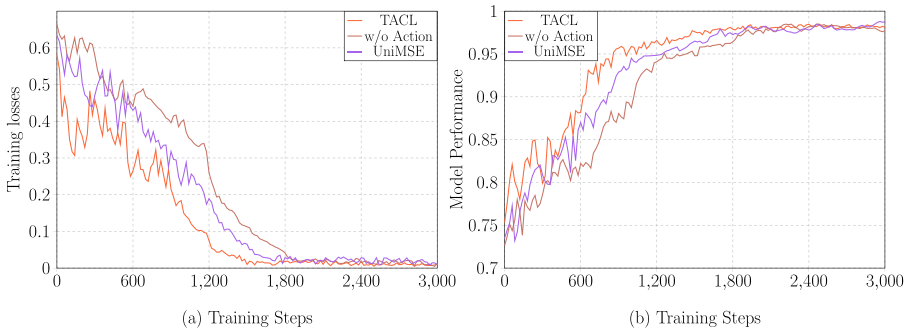



**Fig. 3.** Convergence analysis of TACL, UniMSE, and TACL without Action information.

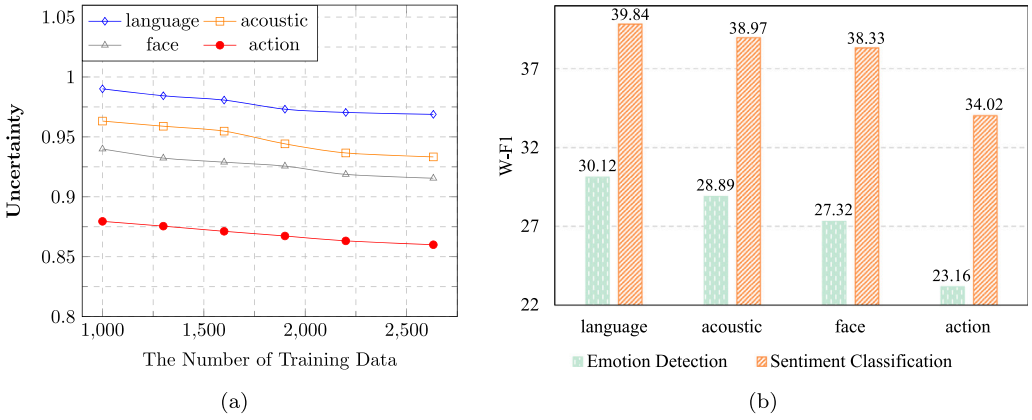



**Fig. 4.** (a) is the uncertainty comparison of the four modalities and the lower uncertainty means the higher credibility. (b) is the single-modality performance comparison.

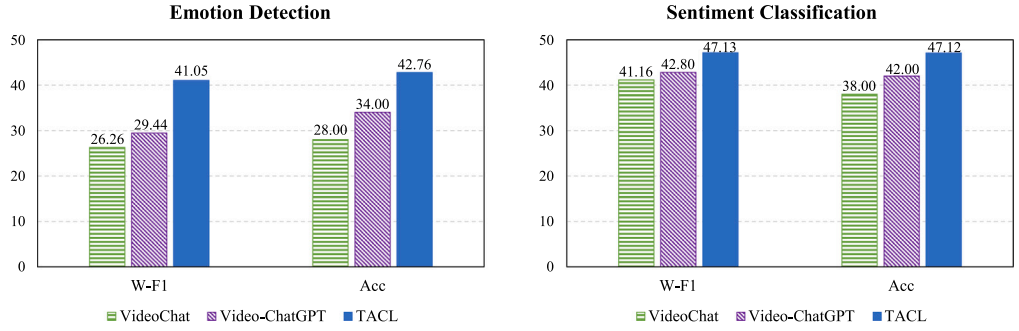



**Fig. 5.** Performance comparison between multimodal video LLMs (i.e., VideoChat, Video-ChatGPT) and our TACL approach in both Emotion Detection and Sentiment Classification tasks.

### 5.4. Comparison with multimodal LLMs

Recently, multimodal large language models (LLMs) have demonstrated their exceptional capabilities in video understanding. To further justify the effectiveness of our proposed TACL approach in MAC, as shown in Fig. 5, we select two open-source multimodal video LLMs (i.e., VideoChat [52] and Video-ChatGPT [53]) to evaluate their performance on the AEV dataset. Specifically, we first randomly select 10 samples for each emotion category (i.e., 50 samples total). Then, we input the videos into the multimodal LLMs along with the prompts, such as "*What kind of emotion does this person express in the video? Choose one from anger, disgust, happy, neutral, and sad.*" and "*What kind of sentiment does this person express in the video? Choose one from positive, neutral, and negative.*" to obtain emotion and sentiment predictions. From Fig. 5, we can observe that TACL outperforms the multimodal LLMs by a large margin, indicating that multimodal LLMs still face challenges in understanding deep affective semantic information in the video.

### 5.5. Visualization and qualitative study

To further justify the effectiveness of our TACL approach for MAC, we provide a visualization and qualitative analysis as shown in Fig. 6. In this case, the man was talking about his personal growth experience. From the language and the loud and energetic tone in the acoustic, we can infer that he feels *proud* and *happy*. And from the facial expressions, we can observe that the man has his mouth wide open in the third frame extracted from the video, which is a facial expression of *surprise* and *happiness*. Therefore, combining information from these three modalities(i.e., language, acoustic, face), the model may tend to predict the emotion of this sample as *happy*.

However, after considering the action information, the predictions of the models tend to be *anger*, which is consistent with the true emotional label. This is because, in the third frame of the action modality, the man makes actions of tilting his head and raising his arm, which is typically an expression of *anger*. Our TACL approach makes correct predictions, whereas UniMSE still maintains incorrect predictions after

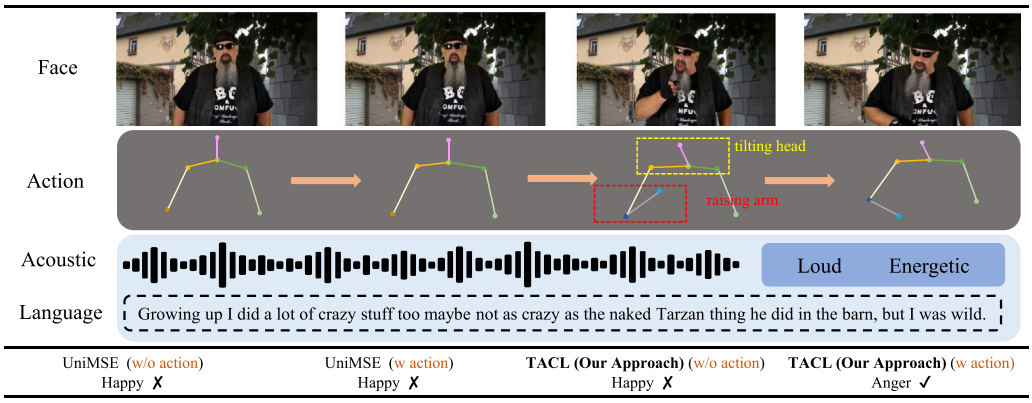



**Fig. 6.** An example from the AEV dataset with its emotion predicted by UniMSE and TACL, where w and w/o action denote incorporating the action information or not, while ✓ (or ✗) denotes that the predicted emotion category is correct (or wrong).

integrating action information, which indicates that TACL is able to address the credibility and sparsity challenges in capturing the action information compared to UniMSE. In addition, the difference of TACL in prediction results between with and without action information also demonstrates the crucial role that action information plays in MAC tasks.

## 6. Conclusion

In this paper, we propose a TACL approach to tackle the credibility and sparsity challenges of the action information for boosting the performance of MAC. Additionally, we construct a high-quality action-enhanced video dataset to evaluate our TACL approach and experimental results on this dataset demonstrate the superior performance of TACL over the state-of-the-art baseline. Furthermore, we observe an interesting finding that the action information is more conducive to facilitating the recognition of negative emotions. In our future work, on one hand, we would like to introduce more information (e.g., micro-expression) to improve the performance of MAC. On the other hand, we intend to transfer our approach to other multimodal tasks, such as multimodal depression detection of which the action information is also crucial.

## CRediT authorship contribution statement

**Tan Yu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jingjing Wang:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Jiamin Luo:** Writing – review & editing, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Jiawen Wang:** Validation, Supervision, Software, Investigation, Formal analysis, Data curation. **Guodong Zhou:** Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by three NSFC grants, i.e., No. 62006166, No. 62376178 and No. 62076175. This work was also supported by the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), China.

## Data availability

Data will be made available on request.

## References

[1] M. Chanchal, B.V. Kumar, Progress in multimodal affective computing: From machine learning to deep learning, in: Proceedings of EAISICC 2023, 2023, pp. 127–150, http://dx.doi.org/10.1007/978-3-031-20541-5_6.

[2] M. Al-Ma'aitah, A. Alwadain, A. Saad, Application dependable interaction module for computer vision-based human-computer interactions, Comput. Electr. Eng. 97 (2022) 107553, http://dx.doi.org/10.1016/j.compeleceng.2021.107553.

[3] R. Kaur, S. Kautish, Multimodal sentiment analysis: A survey and comparison, Int. J. Serv. Sci. Manag. Eng. Technol. 10 (2) (2019) 38–58, http://dx.doi.org/10.4018/IJSSMET.2019040103.

[4] A. Kumar, K. Sharma, A. Sharma, Memor: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries, Image Vis. Comput. 123 (2022) 104483, http://dx.doi.org/10.1016/j.imavis.2022.104483.

[5] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L. Morency, Multi-attention recurrent network for human communication comprehension, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5642–5649, http://dx.doi.org/10.1609/AAAI.V32I1.12024.

[6] Y.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of ACL 2019, 2019, pp. 6558–6569, http://dx.doi.org/10.18653/v1/p19-1656.

[7] M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, L. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, 2018, CoRR abs/1802.00924.

[8] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of EMNLP 2021, 2021, pp. 9180–9192, http://dx.doi.org/10.18653/v1/2021.emnlp-main.723.

[9] M. Liu, K. Liang, Y. Zhao, W. Tu, S. Zhou, X. Gan, X. Liu, K. He, Self-supervised temporal graph learning with temporal and structural intensity alignment, IEEE Trans. Neural Netw. Learn. Syst. (2024).

[10] X. Gao, J. Wang, S. Li, M. Zhang, G. Zhou, Cognition-driven multimodal personality classification, Sci. China Inf. Sci. 65 (10) (2022) http://dx.doi.org/10.1007/S11432-020-3307-3.

[11] T. Yu, J. Wang, J. Wang, J. Luo, G. Zhou, Towards emotion-enriched text-to-motion generation via LLM-guided limb-level emotion manipulating, in: Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024, 2024, http://dx.doi.org/10.1145/3664647.3681487.

[12] Q. Qiao, Y. Xie, J. Gao, T. Wu, S. Huang, J. Fan, Z. Cao, Z. Wang, Y. Zhang, Dntextspotter: arbitrary-shaped scene text spotting via improved denoising training, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 10134–10143.

[13] J. Gao, Q. Qiao, Z. Cao, Z. Wang, W. Li, Aim: let any multi-modal large language models embrace efficient in-context learning, arXiv preprint arXiv:2406.07588 (2024).

[14] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, L. Morency, MTAG: modal-temporal attention graph for unaligned human multimodal language sequences, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of NAACL-HLT, Association for Computational Linguistics, 2021, pp. 1009–1021, http://dx.doi.org/10.18653/V1/2021.NAACL-MAIN.79.

[15] J. Wang, J. Wang, C. Sun, S. Li, X. Liu, L. Si, M. Zhang, G. Zhou, Sentiment classification in customer service dialogue with topic-aware multi-task learning, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, 2020, http://dx.doi.org/10.1609/AAAI.V34I05.6454.

[16] H. Pham, P.P. Liang, T. Manzini, L. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of AAAI 2019, 2019, pp. 6892–6899, http://dx.doi.org/10.1609/aaai.v33i01.33016892.

[17] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of AAAI 2021, 2021, pp. 10790–10797, URL https://ojs.aaai.org/index.php/AAAI/article/view/17289.

[18] J. Wang, C. Sun, S. Li, X. Liu, L. Si, M. Zhang, G. Zhou, Aspect sentiment classification towards question-answering with reinforced bidirectional attention network, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, 2019, http://dx.doi.org/10.18653/V1/P19-1345.

[19] X. Chen, C. Sun, J. Wang, S. Li, L. Si, M. Zhang, G. Zhou, Aspect sentiment classification with document-level sentiment preference modeling, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, 2020, http://dx.doi.org/10.18653/V1/2020.ACL-MAIN.338.

[20] M. Liu, K. Liang, D. Hu, H. Yu, Y. Liu, L. Meng, W. Tu, S. Zhou, X. Liu, Tmac: Temporal multi-modal graph learning for acoustic event classification, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3365–3374.

[21] G. Hu, T. Lin, Y. Zhao, G. Lu, Y. Wu, Y. Li, Unimse: Towards unified multimodal sentiment analysis and emotion recognition, in: Proceedings of EMNLP 2022, 2022, pp. 7837–7851, URL https://aclanthology.org/2022.emnlp-main.534.

[22] J. Wu, Y. Zhang, X. Zhao, W. Gao, A generalized zero-shot framework for emotion recognition from body gestures, 2020, CoRR abs/2010.06362.

[23] S. Baloch, S.A.R.S.A. Bakar, M.M. Mokji, S. Waseem, A. Hafeezallah, Affect recognition using simplistic 2D skeletal features from the upper body movement, in: Proceedings of AICCC 2022, 2022, pp. 96–102, http://dx.doi.org/10.1145/3582099.3582115.

[24] F. Ahmed, A.S.M.H. Bari, M.L. Gavrilova, Emotion recognition from body movement, IEEE Access 8 (2020) 11761–11781, http://dx.doi.org/10.1109/ACCESS.2019.2963113.

[25] Y. Yin, L. Jing, F. Huang, G. Yang, Z. Wang, MSA-GCN: multiscale adaptive graph convolution network for gait emotion recognition, Pattern Recognit. 147 (2024) 110117, http://dx.doi.org/10.1016/J.PATCOG.2023.110117.

[26] E. Marinoiu, M. Zanfir, V. Olaru, C. Sminchisescu, 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism, in: Proceedings of CVPR 2018, 2018, pp. 2158–2167, http://dx.doi.org/10.1109/CVPR.2018.00230.

[27] C. Fantoni, S. Rigutti, W. Gerbino, Bodily action penetrates affective perception, PeerJ 4 (2016) e1677.

[28] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of ICML 2019, 2009, pp. 41–48, http://dx.doi.org/10.1145/1553374.1553380.

[29] Z. Han, C. Zhang, H. Fu, J.T. Zhou, Trusted multi-view classification with dynamic evidential fusion, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2023) 2551–2566, http://dx.doi.org/10.1109/TPAMI.2022.3171983.

[30] A. Graves, M.G. Bellemare, J. Menick, R. Munos, K. Kavukcuoglu, Automated curriculum learning for neural networks, in: Proceedings of ICML 2017, 2017, pp. 1311–1320, URL http://proceedings.mlr.press/v70/graves17a.html.

[31] S. Mai, Y. Sun, H. Hu, Curriculum learning meets weakly supervised multimodal correlation learning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3191–3203.

[32] F. Liu, S. Ge, Y. Zou, X. Wu, Competence-based multimodal curriculum learning for medical report generation, 2022, arXiv preprint arXiv:2206.14579.

[33] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, Librosa: Audio and music signal analysis in python, in: Proceedings of SciPy 2015, 2015, pp. 18–24, http://dx.doi.org/10.25080/Majora-7b98e3ed-003.

[34] T. Baltrusaitis, P. Robinson, L. Morency, OpenFace: An open source facial behavior analysis toolkit, in: Proceedings of WACV 2016, 2016, pp. 1–10, http://dx.doi.org/10.1109/WACV.2016.7477553.

[35] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of CVPR 2019, 2019, pp. 5693–5703, http://dx.doi.org/10.1109/CVPR.2019.00584.

[36] M. Sensoy, L.M. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Proceedings of NeurIPS 2018, 2018, pp. 3183–3193.

[37] A.P. Dempster, A generalization of Bayesian inference, in: R.R. Yager, L. Liu (Eds.), Classic Works of the Dempster-Shafer Theory of Belief Functions, in: Studies in Fuzziness and Soft Computing, vol. 219, Springer, 2008, pp. 73–104, http://dx.doi.org/10.1007/978-3-540-44792-4_4.

[38] A. Jøsang, Subjective logic - A formalism for reasoning under uncertainty, in: Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, 2016, http://dx.doi.org/10.1007/978-3-319-42337-1.

[39] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, vol. 4, (4) Springer, 2006.

[40] E.A. Platanios, O. Stretcu, G. Neubig, B. Póczos, T.M. Mitchell, Competence-based curriculum learning for neural machine translation, in: Proceedings of NAACL 2019, 2019, pp. 1162–1172, http://dx.doi.org/10.18653/v1/n19-1119.

[41] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.

[42] A. Zadeh, R. Zellers, E. Pincus, L. Morency, MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, CoRR abs/1606.06259.

[43] A. Zadeh, P.P. Liang, S. Poria, E. Cambria, L. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2236–2246, http://dx.doi.org/10.18653/v1/P18-1208, URL https://aclanthology.org/P18-1208/.

[44] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 3718–3727, http://dx.doi.org/10.18653/v1/2020.acl-main.343.

[45] P.V.A. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, S. Wermter, The OMG-emotion behavior dataset, in: Proceedings of IJCNN 2018, 2018, pp. 1–7, http://dx.doi.org/10.1109/IJCNN.2018.8489099.

[46] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition, in: Proceedings of ICCV 2013, 2013, pp. 3192–3199, http://dx.doi.org/10.1109/ICCV.2013.396.

[47] J. Cheng, I. Fostiropoulos, B.W. Boehm, M. Soleymani, Multimodal phased transformer for sentiment analysis, in: Proceedings of EMNLP 2021, 2021, pp. 2447–2458, http://dx.doi.org/10.18653/v1/2021.emnlp-main.189.

[48] W. Rahman, M.K. Hasan, S. Lee, A.B. Zadeh, C. Mao, L. Morency, M.E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of ACL 2020, 2020, pp. 2359–2369, http://dx.doi.org/10.18653/v1/2020.acl-main.214.

[49] D. Hazarika, R. Zimmermann, S. Poria, MISA: modality-invariant and -specific representations for multimodal sentiment analysis, in: Proceedings of MM 2020, 2020, pp. 1122–1131, http://dx.doi.org/10.1145/3394171.3413678.

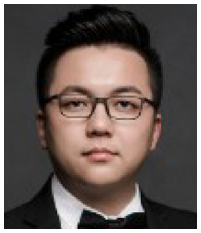
[50] Y. Yang, X. Liu, A re-examination of text categorization methods, in: Proceedings of SIGIR 1999, 1999, pp. 42–49, http://dx.doi.org/10.1145/312624.312647.

[51] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, F. Zheng, Transferable decoding with visual entities for zero-shot image captioning, in: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 3113–3123, http://dx.doi.org/10.1109/ICCV51070.2023.00291.

[52] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, Y. Qiao, VideoChat: Chat-centric video understanding, 2023, http://dx.doi.org/10.48550/ARXIV.2305.06355, CoRR abs/2305.06355.

[53] M. Maaz, H.A. Rasheed, S.H. Khan, F.S. Khan, Video-ChatGPT: Towards detailed video understanding via large vision and language models, 2023, http://dx.doi.org/10.48550/ARXIV.2306.05424, CoRR abs/2306.05424.

**Tan Yu** received the B.E. degree from Soochow University, Suzhou, China. He is currently pursuing his master's degree at Soochow University. His research interests focus on multimodal affective computing and natural language processing.



**JingJing Wang** received the B.E., M.S., and Ph.D. degrees from Soochow University, Suzhou, China. He is currently an Associate Professor at the School of Computer Science and Technology at Soochow University. He serves as an Area Chair and Program Committee member for top international academic conferences in the field of artificial intelligence/natural language processing, such as ACL, EMNLP, COLING, and AAAI. His research interests focus on multimodal affective computing and natural language processing.

**Jiamin Luo** received the B.E. degree from North University of China, Taiyuan, China. She is currently pursuing her Ph.D. at Soochow University. Her research interests focus on multimodal affective computing and natural language processing.

**Jiawen Wang** received the B.E. degree from Soochow University, Suzhou, China. He is currently pursuing his master's degree at Soochow University. His research interests focus on natural language processing.

**Guodong Zhou** graduated with a Ph.D. from the National University of Singapore in December 1997. He has served on the editorial board of the top SCI journal in the international natural language understanding field, Computational Linguistics, and is currently the associate editor of ACM TALLIP and the responsible editor of the Journal of Software. His research areas include natural language understanding, information extraction, and natural language cognition.