Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences

Dingkang Yang [a], Yang Liu [a], Can Huang [b], Mingcheng Li [a], Xiao Zhao [a], Yuzheng Wang [a], Kun Yang [a], Yan Wang [a], Peng Zhai [a,*], Lihua Zhang [a,c,d,*]

[a] *Academy for Engineering and Technology, Fudan University, Shanghai, China*
[b] *School of Journalism, Fudan University, Shanghai, China*
[c] *Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China*
[d] *Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China*

ARTICLE INFO

Article history:
Received 1 December 2022
Received in revised form 2 February 2023
Accepted 3 February 2023
Available online 8 February 2023

Keywords:
Emotion understanding
Knowledge exchange
Multimodal fusion
Crossmodal interaction
Modality co-reinforcement

ABSTRACT

Perceiving human emotions from a multimodal perspective has received significant attention in knowledge engineering communities. Due to the variable receiving frequency for sequences from various modalities, multimodal streams usually have an inherent asynchronous challenge. Most previous methods performed manual sequence alignment before multimodal fusion, which ignored long-range dependencies among modalities and failed to learn reliable crossmodal element correlations. Inspired by the human perception paradigm, we propose a target and source Modality Co-Reinforcement (MCR) approach to achieve sufficient crossmodal interaction and fusion at different granularities. Specifically, MCR introduces two types of target modality reinforcement units to reinforce the multimodal representations jointly. These target units effectively enhance emotion-related knowledge exchange in fine-grained interactions and capture the crossmodal elements that are emotionally expressive in mixed-grained interactions. Moreover, a source modality update module is presented to provide meaningful features for the crossmodal fusion of target modalities. Eventually, the multimodal representations are progressively reinforced and improved via the above components. Comprehensive experiments are conducted on three multimodal emotion understanding benchmarks. Quantitative results show that MCR significantly outperforms the previous state-of-the-art methods in both word-aligned and unaligned settings. Additionally, qualitative analysis and visualization fully demonstrate the superiority of the proposed modules.

1. Introduction

Diverse modalities of human expression provide rich information and knowledge for understanding emotions [1–10]. For multimodal emotion understanding tasks that consider linguistic (*e.g.*, language [11,12]) and non-linguistic (*e.g.*, video and audio [13,14]) modalities, we usually need to perform multimodal fusion across time-series data to extract emotion-related representations. However, the asynchrony among modalities tends to increase the difficulty of analyzing emotions. The collected multimodal streams are unaligned because the sampling rate is variable for the sequences of different modalities. For instance, the video frame with a happy facial expression may relate to a positive word spoken in the past. Exploring the long-term dependencies of asynchronous sequences is a key challenge for effective multimodal fusion.

A straightforward way to overcome the asynchrony challenge is to consider data alignment preceding the fusion. It is the process of finding temporal relations among modalities. For example, most previous works [15–20] performed word-level alignment of textual words via manually pre-processing visual and acoustic sequences, followed by multimodal fusion. These methods would then model the multimodal information on the aligned time steps. Nevertheless, manual alignment operation requires the intervention of feature engineering [19] and consumes a great amount of human resources [21]. Moreover, word-level multimodal fusion ignores the contextual dependencies between elements from different modalities. For direct asynchronous sequence fusion, early work [7] explored the relatedness between multimodal elements according to the maximum mutual information criterion. However, its performance is far from satisfactory due to the shallow learning architecture. Recently, several works [12,22–24] attempted to learn refined multimodal representation based on crossmodal attention. The core insight is

[*] Corresponding authors.
E-mail addresses: dkyang20@fudan.edu.cn (D. Yang), pzhai@fudan.edu.cn (P. Zhai), lihuazhang@fudan.edu.cn (L. Zhang).

https://doi.org/10.1016/j.knosys.2023.110370
0950-7051/© 2023 Elsevier B.V. All rights reserved.

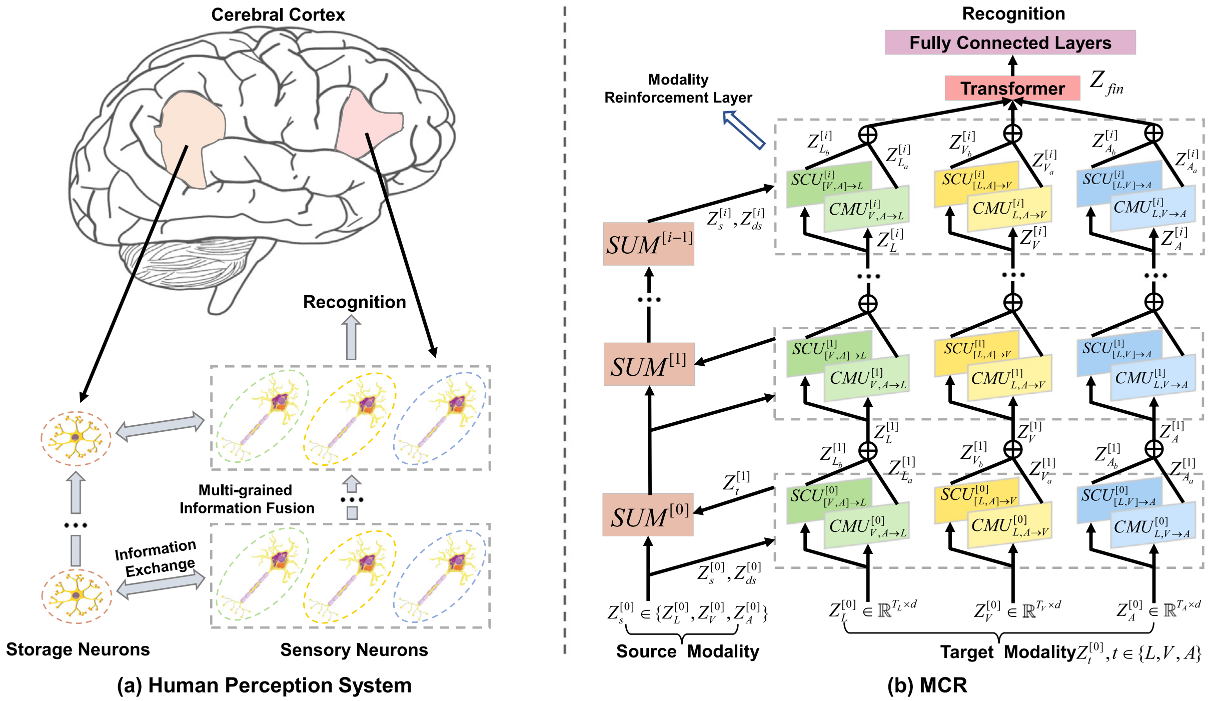

Fig. 1. (a) Illustration of a human multimodal information fusion and perception scheme in the cerebral cortex. (b) Illustration of the proposed MCR. $CMU^{[i]}_{s1,s2\to t}$ allows each source modality jointly to enhance target modalities via fine-grained interactions. $SCU^{[i]}_{[s1,s2]\to t}$ reinforces target modalities via mixed-grained interactions with the concatenated source modalities. Moreover, $SUM^{[i]}$ provides meaningful high-level information of each source modality to reinforce both parts of target modalities in $CMU^{[i+1]}_{s1,s2\to t}$ and $SCU^{[i+1]}_{[s1,s2]\to t}$.

that the transformer-based architecture [25] can elegantly handle asynchronous sequential data and model long-term dependencies among multimodal elements. Unfortunately, they either considered inadequate pairwise interactions between independent modalities [22,24] or focused on coarse global interactions [23]. These deficiencies result in captured crossmodal correlations that could be ambiguous and unreliable.

Recalling whether the superior multimodal fusion procedure in the human perception system has advantages worthy of our consideration. Brain cognitive science studies [26,27] have demonstrated that sensory signals from different organs are sent to the cerebral cortex for unified processing and integration. As shown in Fig. 1(a), the sensory neurons perform signal encoding and fusion of the received multimodal information in parallel patterns at different granularities [28]. During the information exchange at synapses, the storage neurons are employed to retain different levels of multisensory spike trains and update information with the sensory neurons. These specific neurons progressively aggregate multimodal knowledge to achieve accurate human perception [29].

Inspired by the above multisensory fusion scheme, we propose a target and source Modality Co-Reinforcement (MCR) approach for multimodal fusion from asynchronous multimodal sequences. The architecture of MCR is illustrated in Fig. 1(b). Imitating the multi-grained information fusion of sensory neurons, MCR introduces two types of parallel reinforcement units to jointly reinforce target modalities' representations. These elaborate units include the Cross-Mutual attention reinforcement Unit (*CMU*) and Self-Cross attention reinforcement Unit (*SCU*). On the one hand, *CMU* elegantly integrates fine-grained crossmodal interactions between target modalities and each source modality to enhance information and knowledge exchange. In this case, a mutual attention strategy is presented to learn the complementarity of crossmodal attention. On the other hand, *SCU* focuses on capturing crossmodal elements with stronger correlations in mixed-grained interactions between target modalities and integrated source modalities to learn reliable contextual dependencies. Concurrently, mimicking the information exchange of storage neurons, we propose an accompanying Source modality Update Module (*SUM*) to reinforce source modalities progressively. *SUM* updates the specific representation of each source modality independently and under supervision by aggregating the reinforced features of the corresponding target modality from the next layer. Meanwhile, a selective memory mechanism is applied for information filtering to provide valuable high-level features for *CMU* and *SCU* that perform crossmodal interactions. Therefore, the features of target and source modalities are improved to obtain the final robust representations for downstream emotion understanding tasks. The main contributions can be summarized as follows:

- We propose MCR, a human perception paradigm-driven approach to directly address asynchronous multimodal sequence fusion and interaction.
- We introduce two types of target modality reinforcement units to capture the emotion-related correlations of crossmodal elements at different granularities. Additionally, a source modality update module is proposed to progressively provide effective high-level features for crossmodal interactions with target modalities.
- We evaluate MCR on three multimodal emotion understanding benchmarks, including MOSI [19], MOSEI [30], and IEMOCAP [31]. Comprehensive experiments show that MCR outperforms existing state-of-the-art methods in both word-aligned and unaligned settings.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 details the proposed approach and its modules. Section 4 describes the experimental settings, results, and visualizations. Section 5 concludes this paper.

2. Related work

This section reviews works in two aspects: video-based multimodal emotion understanding and transformer network.

2.1. Video-based multimodal emotion understanding

Video flows usually involve time-series data from multiple modalities, such as natural language, visual information, and acoustic behaviors. Analyzing videos from a multimodal perspective can facilitate a superior understanding of human emotions [32–36]. Unlike learning multimodal representations from static domains such as image and text attributes [37–42], multimodal time-series data focuses on learning long-term dependencies among modalities. A core point is combining diverse modalities to perform effective multimodal fusion. Early fusion methods [43–45] considered each modality to have an equal contribution and performed simple feature concatenation. In addition, several works [5,46–49] attempted to utilize intermediate and late fusion strategies that combine the specific representation learned by each modality. Despite the significant improvements achieved compared to considering a single modality, they failed to explicitly capture the inherent dependencies among different modalities.

In practice, the multimodal streams are usually asynchronous due to the variable frame rate for sequences of different modalities. To tackle this issue, existing methods [8,13,15–18,20,50–52] manually aligned visual and acoustic sequences to the textual words before performing multimodal fusion. As examples of early efforts, RAVEN [15] dynamically shifted word representations based on nonverbal cues and applied the attention-gating module to fuse the word-level features. MCTN [16] enforced translation from a source modality to a target modality, resulting in an intermediate representation that captures joint information. For recent works, TCSP [18] designed a crossmodal prediction task to explore the shared and private semantics for each non-textual modality. FDMER [8] focused on distilling distinct representations for each modality via a feature-disentangled strategy. However, manual alignment usually requires extensive labor efforts and ignores dependencies among multimodal elements.

For directly dealing with asynchronous multimodal sequences, MulT [22] is proposed to learn the potential adaption of a modality to another, thus repeatedly reinforcing target modalities. Nevertheless, MulT only considered interactions between two modalities, and the interaction pairs in each direction did not exchange information. After that, PMR [23] improved MulT by introducing a message hub to obtain the reinforced features of source modalities. However, the concatenation of all modalities in the message hub destroyed the specific representation of each source modality. More recent works [12,24] advocated learning correlations among elements in the modality-specific or -agnostic space but failed to consider the importance of comprehensive interaction in multimodal fusion. In contrast, the proposed approach can progressively learn correlations between elements of multimodal sequences via multi-grained crossmodal interactions. Benefiting from our stacking strategy, we achieve multi-level information exchange of different modalities.

2.2. Transformer network

Transformer [25] is the state-of-the-art for sequential modeling which achieves superior performances. Compared to the recurrent neural network [53] and long short-term memory network [54], the self-attention operation of the transformer is more effective in exploring the element correlations. The network architecture is first introduced for machine translation task [55], where the encoder and decoder use self-attention transformers. In addition to language-related applications, the transformer has also been adopted in computer vision [56,57], audio processing [58,59] and even other disciplines [60,61].

Unlike the vanilla encoder–decoder structure, our MCR is built over tailored modality reinforcement units following transformer-like encoder structures. we extend the vanilla attention to the crossmodal attention paradigm by introducing the query and key vectors from distinct modalities. The core crossmodal attention is powerful for learning long-term dependencies and refining feature representations between elements across modalities. Moreover, we extend the conventional unimodal-based transformer encoder to the multimodal paradigm. The proposed modality co-reinforcement components provide the potential to achieve a unified multimodal fusion procedure in different sequence settings.

3. Methodology

In this section, we first provide the problem statement and preliminary. Then, we show an overview of the proposed approach. After that, we elaborate separately on the two types of target modality reinforcement units and the source modality update module.

3.1. Problem statement and preliminary

3.1.1. Problem statement

Based on the diversity of emotional expression, this work considers three primary modalities, *i.e.*, language (L), video (V), and audio (A) modalities. Given three sequences from these modalities as $\boldsymbol{X}_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$. $T_{\{\cdot\}}$ and $d_{\{\cdot\}}$ represent sequence length and feature dimension, respectively. We aim to perform effective multimodal fusion and interaction from unaligned multimodal sequences. Then, the reinforced multimodal representation is used to implement different emotion understanding tasks.

3.1.2. Crossmodal attention

The core idea of crossmodal attention is to repeatedly reinforce the representation of a target modality via directional pairwise interactions across target and source modalities [22]. Define the sequence of the target modality as $\boldsymbol{X}_t \in \mathbb{R}^{T_t \times d_t}$ and the sequence of the source modality as $\boldsymbol{X}_s \in \mathbb{R}^{T_s \times d_s}$, where $t, s \in \{L, V, A\}$. Inspired by the self-attention [25], the crossmodal attention contains Querys, Keys, and Values, denoted as $\boldsymbol{Q}_t = \boldsymbol{X}_t \boldsymbol{W}_{Q_t}$ with $\boldsymbol{W}_{Q_t} \in \mathbb{R}^{d_t \times d_k}$, $\boldsymbol{K}_s = \boldsymbol{X}_s \boldsymbol{W}_{K_s}$ with $\boldsymbol{W}_{K_s} \in \mathbb{R}^{d_s \times d_k}$ and $\boldsymbol{V}_s = \boldsymbol{X}_s \boldsymbol{W}_{V_s}$ with $\boldsymbol{W}_{V_s} \in \mathbb{R}^{d_s \times d_v}$, respectively. One single head of the crossmodal attention can be formulated as:

$$\begin{aligned} \boldsymbol{Y}_t &= CA_{s \to t}(\boldsymbol{X}_s, \boldsymbol{X}_t), \\ &= softmax(\frac{\boldsymbol{Q}_t \boldsymbol{K}_s^T}{\sqrt{d_k}})\boldsymbol{V}_s, \end{aligned} \tag{1}$$

where $\boldsymbol{Y}_t \in \mathbb{R}^{T_t \times d_v}$. The h-head crossmodal attention is denoted as $\boldsymbol{Y}_t = CA^{mul}_{s \to t}(\boldsymbol{X}_s, \boldsymbol{X}_t)$, where $\boldsymbol{Y}_t \in \mathbb{R}^{T_t \times hd_v}$.

3.2. Model overview

First, a 1D temporal convolutional layer is used to improve the neighborhood perception of the multimodal sequences. By controlling the kernel size of the convolutional operation for each modality, the features of different modalities are enforced to have identical dimensions. Following [25], the positional embedding is added to the input sequences to complete the preprocessing. The pre-processed sequences are denoted as $\boldsymbol{Z}_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$. As shown in Fig. 1(b), MCR reinforces each target modality in concert with two types of parallel reinforcement units

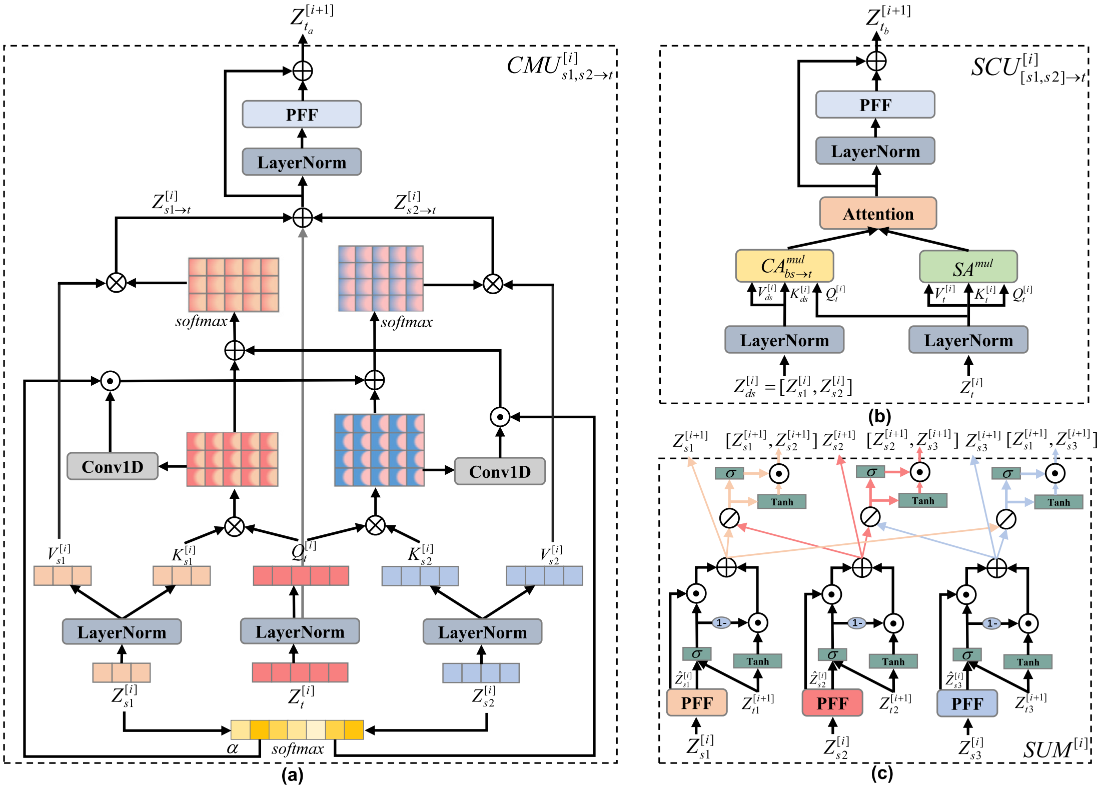
$Z_{t_a}^{[i+1]}$
$CMU_{s1,s2\rightarrow t}^{[i]}$
PFF
LayerNorm
$Z_{s1\rightarrow t}^{[i]}$
$Z_{s2\rightarrow t}^{[i]}$
softmax
softmax
Conv1D
Conv1D
$V_{s1}^{[i]}$
$K_{s1}^{[i]}$
$Q_t^{[i]}$
$K_{s2}^{[i]}$
$V_{s2}^{[i]}$
LayerNorm
LayerNorm
LayerNorm
$Z_{s1}^{[i]}$
$Z_t^{[i]}$
$Z_{s2}^{[i]}$
α
softmax
(a)
$Z_{t_b}^{[i+1]}$
$SCU_{[s1,s2]\rightarrow t}^{[i]}$
PFF
LayerNorm
Attention
$CA_{ds\rightarrow t}^{mul}$
SA^{mul}
$V_{ds}^{[i]}$ $K_{ds}^{[i]}$ $Q_t^{[i]}$
$V_t^{[i]}$ $K_t^{[i]}$ $Q_t^{[i]}$
LayerNorm
LayerNorm
$Z_{ds}^{[i]}=[Z_{s1}^{[i]},Z_{s2}^{[i]}]$
$Z_t^{[i]}$
(b)
$Z_{s1}^{[i+1]}$ $[Z_{s1}^{[i+1]},Z_{s2}^{[i+1]}]$ $Z_{s2}^{[i+1]}$ $[Z_{s2}^{[i+1]},Z_{s3}^{[i+1]}]$ $Z_{s3}^{[i+1]}$ $[Z_{s1}^{[i+1]},Z_{s3}^{[i+1]}]$
σ
Tanh
1-
$\hat{Z}_{s1}^{[i]}$ $\hat{Z}_{s2}^{[i]}$ $\hat{Z}_{s3}^{[i]}$
PFF
PFF
PFF
$Z_{t1}^{[i+1]}$ $Z_{t2}^{[i+1]}$ $Z_{t3}^{[i+1]}$
$Z_{s1}^{[i]}$ $Z_{s2}^{[i]}$ $Z_{s3}^{[i]}$
$SUM^{[i]}$
(c)

Fig. 2. The architecture of the proposed (a) $CMU_{s1,s2\rightarrow t}^{[i]}$, (b) $SCU_{[s1,s2]\rightarrow t}^{[i]}$, and (c) $SUM^{[i]}$. PFF represents the position-wise feed-forward layer. ⊙ denotes the Hadamard product. ⊕ denotes the element-wise sum. ⊗ denotes the matrix multiplication. ⊘ denotes the concatenation operation. σ denotes the sigmoid activation.

(*i.e.*, $CMU_{s1,s2\rightarrow t}$ and $SCU_{[s1,s2]\rightarrow t}$). Specifically, $CMU_{s1,s2\rightarrow t}$ obtains the first-partially reinforced target modality $\boldsymbol{Z}_{t_a} \in \mathbb{R}^{T_t\times d}$ via the source modalities $\boldsymbol{Z}_{s1} \in \mathbb{R}^{T_{s1}\times d}$ and $\boldsymbol{Z}_{s2} \in \mathbb{R}^{T_{s2}\times d}$, where $t, s1, s2 \in \{L, V, A\}$. Concurrently, $SCU_{[s1,s2]\rightarrow t}$ obtains the second-partially reinforced target modality $\boldsymbol{Z}_{t_b} \in \mathbb{R}^{T_t\times d}$ via concatenating the source modalities involved in the corresponding $CMU_{s1,s2\rightarrow t}$ as $\boldsymbol{Z}_{ds} = [\boldsymbol{Z}_{s1}, \boldsymbol{Z}_{s2}] \in \mathbb{R}^{(T_{s1}+T_{s2})\times d}$. Immediately, the outputs from the modality-specific $CMU_{s1,s2\rightarrow t}$ and $SCU_{[s1,s2]\rightarrow t}$ are merged to obtain the reinforced representations of target modalities as $\boldsymbol{Z}_t = (\boldsymbol{Z}_{t_a} + \boldsymbol{Z}_{t_b}) \in \mathbb{R}^{T_t\times d}$, where $t \in \{L, V, A\}$. Additionally, the accompanying *SUM* updates each source modality to provide reinforced high-level features for crossmodal fusion and interaction. Our core structure is a multi-layer stacking of the above components to progressively achieve co-reinforcement of target and source modalities. We then concatenate all the reinforced features of target modalities as $\boldsymbol{Z}_{fin} = [\boldsymbol{Z}_L, \boldsymbol{Z}_V, \boldsymbol{Z}_A] \in \mathbb{R}^{(T_L+T_V+T_A)\times d}$ fed to a transformer encoder [25] to collect temporal information. Finally, the fully connected layers are utilized for emotion recognition or sentiment analysis.

3.3. Target modality reinforcement

By imitating the working pattern of multisensory neurons in the cerebral cortex [28], we achieve multimodal fusion with different granularities of crossmodal interactions to capture reliable element correlations among multimodal sequences. Concretely, $CMU_{s1,s2\rightarrow t}$ focuses on fine-grained fusion and interaction among independent modalities, while $SCU_{[s1,s2]\rightarrow t}$ attends to mixed-grained fusion and interaction among integrated modalities. These two target units learn effective multimodal representations progressively in collaboration with each other. The reinforcement of the target modalities is achieved by two parts of the stacked $CMU_{s1,s2\rightarrow t}^{[i]}$ and $SCU_{[s1,s2]\rightarrow t}^{[i]}$ together, where $t, s1, s2 \in \{L, V, A\}, t \neq s1, s2$. The superscript $[i]$ indicates the ith modality reinforcement layer. Note that the pre-processed multimodal sequences $\boldsymbol{Z}_{\{L,V,A\}}$ are used to generate a copy as source modalities. During the co-reinforcement, the source modalities are only updated in $SUM^{[i]}$ to reinforce the target modalities and are not utilized as the final output.

3.3.1. Cross-mutual attention reinforcement unit

The previous method [22] performed pairwise interactions between the target modality and each source modality independently, with only a simple concatenation at the final stage. This pattern leads to limited performance gains [23,24]. In contrast, the proposed $CMU_{s1,s2\rightarrow t}^{[i]}$ allows each crossmodal interaction to cooperate and share information in a single update unit efficiently. Considering that some sequential data of modalities are usually noisy and redundant [62], we also introduce a mutual attention mechanism to improve the reliability of source modalities' attention. The structure of $CMU_{s1,s2\rightarrow t}^{[i]}$ is shown in Fig. 2(a). Its inputs are $\boldsymbol{Z}_{s1}^{[i]}$, $\boldsymbol{Z}_{s2}^{[i]}$ and $\boldsymbol{Z}_t^{[i]}$ while its output is the reinforced features of the target modality as $\boldsymbol{Z}_{t_a}^{[i+1]} \in \mathbb{R}^{T_t\times d}$:

$$\boldsymbol{Z}_{t_a}^{[i+1]} = CMU_{s1,s2\rightarrow t}^{[i]}(\boldsymbol{Z}_{s1}^{[i]}, \boldsymbol{Z}_{s2}^{[i]}, \boldsymbol{Z}_t^{[i]}). \tag{2}$$

Formally, the Keys & Values from the source modalities and the Querys from the target modality are denoted as $\boldsymbol{K}_*^{[i]} = LN(\boldsymbol{Z}_*^{[i]})\boldsymbol{W}_{K_*}^{[i]} \in \mathbb{R}^{T_*\times d}$, $\boldsymbol{V}_*^{[i]} = LN(\boldsymbol{Z}_*^{[i]})\boldsymbol{W}_{V_*}^{[i]} \in \mathbb{R}^{T_*\times d}$, and $\boldsymbol{Q}_t^{[i]} = LN(\boldsymbol{Z}_t^{[i]})\boldsymbol{W}_{Q_t}^{[i]} \in \mathbb{R}^{T_t\times d}$, respectively, where $* \in \{s1, s2\}$. LN means layer normalization. The latent adaptation from a source modality to a target modality is denoted as $\boldsymbol{E}_{*\rightarrow t}^{[i]} = \frac{\boldsymbol{Q}_t^{[i]}\boldsymbol{K}_*^{[i]^T}}{\sqrt{d}} \in \mathbb{R}^{T_t\times T_*}$. Then, both crossmodal interactions exchange dimensions in the direction of the source sequence via a 1D convolutional layer denoted as $\hat{\boldsymbol{E}}_{s1\rightarrow t}^{[i]} \in \mathbb{R}^{T_t\times T_{s2}}$ and $\hat{\boldsymbol{E}}_{s2\rightarrow t}^{[i]} \in \mathbb{R}^{T_t\times T_{s1}}$, respectively. After concatenation, $\boldsymbol{Z}_{s1}^{[i]}$ and $\boldsymbol{Z}_{s2}^{[i]}$ are summed over the feature dimension at each time step, and then activated via a softmax function to obtain selective attention weights for the source modalities:

$$\alpha^{[i]} = softmax([\boldsymbol{Z}_{s1}^{[i]}, \boldsymbol{Z}_{s2}^{[i]}] \cdot I^{[i]}) \in \mathbb{R}^{(T_{s1}+T_{s2})\times 1}, \tag{3}$$

where $I^{[i]} \in \mathbb{R}^{d\times 1}$ is an all-1 vector. $\alpha^{[i]}$ is further divided into $\alpha_{s1}^{[i]} \in \mathbb{R}^{T_{s1}\times 1}$ and $\alpha_{s2}^{[i]} \in \mathbb{R}^{T_{s2}\times 1}$, indicating the reliability of the features on each source modality's sequence. Immediately, the crossmodal fusion combined re-weighted mutual attention is expressed as follows:

$$\boldsymbol{Z}_{s1\to t}^{[i]} = softmax(\boldsymbol{E}_{s1\to t}^{[i]} + \alpha_{s1}^{[i]} \odot \hat{\boldsymbol{E}}_{s2\to t}^{[i]})\boldsymbol{V}_{s1}^{[i]} \in \mathbb{R}^{T_t\times d}, \tag{4}$$

$$\boldsymbol{Z}_{s2\to t}^{[i]} = softmax(\boldsymbol{E}_{s2\to t}^{[i]} + \alpha_{s2}^{[i]} \odot \hat{\boldsymbol{E}}_{s1\to t}^{[i]})\boldsymbol{V}_{s2}^{[i]} \in \mathbb{R}^{T_t\times d}. \tag{5}$$

Finally, the reinforced features of the target modality are merged by element-wise sum operation as $\boldsymbol{Z}_{t_a}^{[i]} = (LN(\boldsymbol{Z}_t^{[i]}) + \boldsymbol{Z}_{s1\to t}^{[i]} + \boldsymbol{Z}_{s2\to t}^{[i]}) \in \mathbb{R}^{T_t\times d}$. As in the transformer model [25], a position-wise feed-forward layer with skip connection will process $\boldsymbol{Z}_{t_a}^{[i]}$ and obtain $\boldsymbol{Z}_{t_a}^{[i+1]}$ for the next modality reinforcement layer.

3.3.2. Self-cross attention reinforcement unit

It is essential in multimodal fusion to perform mixed-grained interactions following the human perception system [27]. $SCU_{[s1,s2]\to t}^{[i]}$ improves the ability to explore the correlations with the target modality under source modalities co-occurrence. The structure of $SCU_{[s1,s2]\to t}^{[i]}$ is shown in Fig. 2(b). Its inputs are $\boldsymbol{Z}_{ds}^{[i]} = [\boldsymbol{Z}_{s1}^{[i]}, \boldsymbol{Z}_{s2}^{[i]}] \in \mathbb{R}^{(T_{s1}+T_{s2})\times d}$ and $\boldsymbol{Z}_t^{[i]}$ while its output is the reinforced features of the target modality as $\boldsymbol{Z}_{t_b}^{[i+1]}$:

$$\boldsymbol{Z}_{t_b}^{[i+1]} = SCU_{[s1,s2]\to t}^{[i]}(\boldsymbol{Z}_{ds}^{[i]}, \boldsymbol{Z}_t^{[i]}), \tag{6}$$

where $\boldsymbol{Z}_{t_b}^{[i+1]} \in \mathbb{R}^{T_t\times d}$. To be specific, $\boldsymbol{Z}_t^{[i]}$ is reinforced via two branches of the self-attention and crossmodal attention:

$$\boldsymbol{Z}_{ds\to t}^{[i]} = CA_{ds\to t}^{mul}(LN(\boldsymbol{Z}_{ds}^{[i]}), LN(\boldsymbol{Z}_t^{[i]})), \tag{7}$$

$$\boldsymbol{Z}_t^{[i]} = SA^{mul}(LN(\boldsymbol{Z}_t^{[i]})), \tag{8}$$

where $\boldsymbol{Z}_{ds\to t}^{[i]}, \boldsymbol{Z}_t^{[i]} \in \mathbb{R}^{T_t\times d}$, and SA^{mul} means multi-head self-attention operation [25]. The fundamental insight introduced by two distinct branches is that capturing both intra-modal dynamics and inter-modal dependencies enhances the target representations. Subsequently, a novel attention layer is utilized to obtain the reinforced feature $\boldsymbol{Z}_{t_b}^{[i]}$ from the adapted fusion. The attention layer can assign dynamic weights for each reinforced representation based on its importance. To this end, $\boldsymbol{Z}_{ds\to t}^{[i]}, \boldsymbol{Z}_t^{[i]}$ are reshaped as $\hat{\boldsymbol{Z}}_{ds\to t}^{[i]}, \hat{\boldsymbol{Z}}_t^{[i]} \in \mathbb{R}^{T_t\cdot d\times 1}$, respectively. The process of the attention layer is as follows:

$$\beta_*^{[i]} = \boldsymbol{U}^T(\boldsymbol{W}_*^{[i]} \cdot \hat{\boldsymbol{Z}}_*^{[i]} + \boldsymbol{b}_*^{[i]}), \tag{9}$$

$$\gamma_*^{[i]} = \frac{exp(\beta_*^{[i]})}{\sum_{*\in\{ds\to t,t\}} exp(\beta_*^{[i]})}, \tag{10}$$

$$\boldsymbol{Z}_{t_b}^{[i]} = \sum_{*\in\{ds\to t,t\}} \gamma_*^{[i]} \odot \boldsymbol{Z}_*^{[i]}, \tag{11}$$

where $\boldsymbol{U} \in \mathbb{R}^{T_t\cdot d\times 1}$, $\boldsymbol{W}_*^{[i]} \in \mathbb{R}^{T_t\cdot d\times T_t\cdot d}$, and $\boldsymbol{b}_*^{[i]} \in \mathbb{R}^{T_t\cdot d\times 1}$ are learnable parameters. Finally, we pass $\boldsymbol{Z}_{t_b}^{[i]}$ through a position-wise feed-forward layer with skip connection and obtain $\boldsymbol{Z}_{t_b}^{[i+1]}$.

3.4. Source modality update module

Most previous works [12,23] aimed at concatenating source modalities to reinforce them. These operations destroyed each modality's specific representation and potentially introduced redundant and noisy information. In comparison, mimicking the information exchange mechanism of the storage neurons [29], $SUM^{[i]}$ can provide reliable high-level features of each source modality for the next layer of $CMU_{s1,s2\to t}^{[i+1]}$ and $SCU_{[s1,s2]\to t}^{[i+1]}$. As shown in Fig. 2(c), each source modality is explicitly supervised and attended by the reinforced features of the corresponding target modality as $\boldsymbol{Z}_t^{[i+1]} = (\boldsymbol{Z}_{t_a}^{[i+1]} + \boldsymbol{Z}_{t_b}^{[i+1]}) \in \mathbb{R}^{T_t\times d}$, where $t \in \{L, V, A\}$. Its inputs are $\boldsymbol{Z}_t^{[i+1]}$ and $\boldsymbol{Z}_s^{[i]}$ and its outputs are the reinforced features $\boldsymbol{Z}_s^{[i+1]}$ and $\boldsymbol{Z}_{ds}^{[i+1]}$:

$$\boldsymbol{Z}_s^{[i+1]}, \boldsymbol{Z}_{ds}^{[i+1]} = SUM^{[i]}(\boldsymbol{Z}_t^{[i+1]}, \boldsymbol{Z}_s^{[i]}), \tag{12}$$

where $s, t \in \{L, V, A\}$. $\boldsymbol{Z}_{ds}^{[i+1]} \in \mathbb{R}^{(T_{s1}+T_{s2})\times d}$ is generated based on the combination of source modalities $\boldsymbol{Z}_s^{[i+1]} \in \mathbb{R}^{T_s\times d}$ required by $SCU_{[s1,s2]\to t}^{[i+1]}$ for different target modalities. More formally, each $\boldsymbol{Z}_s^{[i]}$ first achieves self-updating of the representation via a feed-forward layer, denoted as $\hat{\boldsymbol{Z}}_s^{[i]} = \mathcal{F}_s(\boldsymbol{Z}_s^{[i]}; \boldsymbol{W}_\theta^{[i]}) \in \mathbb{R}^{T_s\times d}$, where $\boldsymbol{W}_\theta^{[i]}$ are network parameters. Subsequently, a selective memory mechanism is proposed to implement information filtering, This process aims to retain the meaningful information in $\hat{\boldsymbol{Z}}_s^{[i]}$ while forgetting the redundant information in the corresponding $\boldsymbol{Z}_t^{[i+1]}$. The formula is expressed as follows:

$$\mu_s^{[i]} = \sigma(\hat{\boldsymbol{Z}}_s^{[i]} \cdot \hat{\boldsymbol{W}}_s^{[i]} + \boldsymbol{Z}_t^{[i+1]} \cdot \boldsymbol{W}_t^{[i+1]} + \hat{\boldsymbol{b}}_s^{[i]} + \boldsymbol{b}_t^{[i+1]}), \tag{13}$$

$$\boldsymbol{Z}_s^{[i+1]} = (1 - \mu_s^{[i]}) \odot tanh(\boldsymbol{Z}_t^{[i+1]}) + \mu_s^{[i]} \odot \hat{\boldsymbol{Z}}_s^{[i]}, \tag{14}$$

where $\hat{\boldsymbol{W}}_s^{[i]}, \boldsymbol{W}_t^{[i+1]} \in \mathbb{R}^{d\times d}$, $\hat{\boldsymbol{b}}_s^{[i]}, \boldsymbol{b}_t^{[i+1]} \in \mathbb{R}^{T_s\times d}$ are the learnable parameters. The reinforced features of each source modality $\boldsymbol{Z}_s^{[i+1]}$ generated by the above mechanism are applied to the corresponding next layer of $CMU_{s1,s2\to t}^{[i+1]}$. In addition, the $\boldsymbol{Z}_s^{[i+1]}$ perform further information updates based on different combinations to obtain well-chosen $\boldsymbol{Z}_{ds}^{[i+1]}$ for $SCU_{[s1,s2]\to t}^{[i+1]}$:

$$\varphi_s^{[i]} = \sigma([\boldsymbol{Z}_{s1}^{[i+1]}, \boldsymbol{Z}_{s2}^{[i+1]}] \cdot \boldsymbol{W}_s^{[i]} + \boldsymbol{b}_s^{[i]}), \tag{15}$$

$$\boldsymbol{Z}_{ds}^{[i+1]} = \varphi_s^{[i]} \odot tanh([\boldsymbol{Z}_{s1}^{[i+1]}, \boldsymbol{Z}_{s2}^{[i+1]}]), \tag{16}$$

where $\boldsymbol{W}_s^{[i]} \in \mathbb{R}^{d\times d}$, $\boldsymbol{b}_s^{[i]} \in \mathbb{R}^{(T_{s1}+T_{s2})\times d}$ are the learnable parameters.

4. Experiments

In this section, we first detail experimental settings, including benchmarks, evaluation metrics, and implementation details. Then, we compare our approach with state-of-the-art methods and conduct visualization analyses. Finally, we perform ablation studies.

4.1. Benchmarks and evaluation metrics

Comprehensive experiments are conducted on three video-based multimodal benchmarks, including two sentiment analysis tasks: MOSI [19] and MOSEI [30], and an emotion recognition task: IEMOCAP [31].

4.1.1. MOSI and MOSEI

MOSI [19] is a dataset containing 2199 short monologue video clips. The standard partitioning of the dataset is 1284 samples in the training set, 229 in the validation set, and 686 in the testing set. Meanwhile, the textual data are segmented per word and expressed as discrete word embeddings. MOSEI [30] contains 22,856 annotated video segments from 1000 distinct speakers and 250 topics acquired from social media channels. Its predetermined data partition has 16,326 samples in the training set, 1871 in the validation set, and 4659 in the testing set. For MOSI and MOSEI, each sample is labeled by human annotators with a sentiment score from −3 of strongly negative to 3 of strongly positive. To thoroughly evaluate our approach, we adopt diverse

Table 1
The hyper-parameter settings in each benchmark.

Setting	MOSI	MOSEI	IEMOCAP
Batch size	64	16	32
Learning rate	1e−3	1e−3	2e−3
Epoch number	120	30	60
Feature size	40	40	40
Attention head	10	8	10
Kernel size ($L/V/A$)	3/3/3	3/3/3	3/3/5
Reinforcement layer	3	4	3

metrics, including Acc_7: 7-class accuracy of emotion score classification, Acc_2: binary accuracy of positive/negative emotions, $F1$ score, MAE: mean absolute error, and $Corr$: the correlation of the model's prediction with human.

4.1.2. IEMOCAP

IEMOCAP [31] consists of text, audio, and video modalities of 10 actors recorded in the form of conversations using a Motion Capture camera. As suggested by [15], 4 emotions (*i.e.*, happy, sad, angry, and neutral) are selected for emotion recognition. Following the well-known previous works [17,63], we report Acc: the binary classification accuracy and $F1$ score for each category.

4.2. Implementation details

We convert video transcripts into pre-trained Glove word embedding [64] with 300-dimensional vectors. For the audio, we employ COVAREP toolkit [65] for extracting 74-dimensional low-level acoustic features. The features include 12 Mel-frequency cepstral coefficients (MFCCs), voiced/unvoiced segmenting features, glottal source parameters, etc. Meanwhile, the Facet [66] is used to indicate 35 facial action units, which record facial muscle movement for representing per-frame emotions. All models are built on the Pytorch toolbox with two Quadro RTX 8000 GPUs. The Adam optimizer [67] is adopted for network optimization. For the classification and regression tasks, we use the standard cross-entropy loss and L_1 loss separately to guarantee a fair comparison [22]. Table 1 shows the hyper-parameter settings adopted for each benchmark. The kernel size is set for the 1D convolutional operation used to preprocess the sequence of each modality. Each modality reinforcement layer consists of two types of reinforcement units (*i.e.*, $CMU_{s1,s2\to t}$ and $SCU_{[s1,s2]\to t}$). The hyper-parameters are determined via the validation set.

4.3. Comparison with state-of-the-art methods

We comprehensively compare our MCR with state-of-the-art (SOTA) methods, including Recurrent Multistage Fusion Network (RMFN) [20], Recurrent Attended Variation Embedding Network (RAVEN) [15], Multimodal Cyclic Translation Network (MCTN) [16], Multimodal Factorization Model (MFM) [17], Graph-MFN [30], Multimodal Transformer (MulT) [22], Text-Centered Shared-Private Network (TCSP) [18], Modal-Temporal Attention Graph (MTAG) [68], Progressive Modality Reinforcement (PMR) [23], Modality-Invariant Crossmodal Attention (MICA) [24], Feature-Disentangled MER (FDMER) [8], and MFSA [12]. For the prominent works [15,16] in the word-aligned setting, we optimize the alignment objective to match the unaligned setting by adding Connectionist Temporal Classification (CTC) loss [21]. Furthermore, we report the average results of all metrics over five experiments with different random seeds.

Table 2
Comparison on the MOSI benchmark. Best results are marked in **bold** and † means the corresponding result is significantly better than SOTA with p-value ¡ 0.05 based on paired t-test. The footnote † to Tables 3 and 4 follow identical interpretations.

Model	Acc_7 ↑	Acc_2 ↑	$F1$ ↑	MAE ↓	$Corr$ ↑
Word-aligned setting					
Graph-MFN [30]	29.6	75.4	75.2	0.963	0.618
RAVEN [15]	33.2	78.0	76.6	0.915	0.691
MCTN [16]	35.6	79.3	79.1	0.909	0.676
MFM [17]	36.2	78.1	78.1	0.951	0.662
RMFN [20]	38.3	78.4	78.0	0.922	0.681
MulT [22]	40.0	83.0	82.8	0.871	0.698
TCSP [18]	–	80.9	81.0	0.908	0.710
PMR [23]	40.6	83.6	83.4	–	–
FDMER [8]	42.1	84.2	83.9	0.845	0.732
MCR (ours)	**43.8**†	**86.5**†	**85.7**†	**0.802**†	**0.747**†
Unaligned setting					
RAVEN [15]	31.7	72.7	73.1	1.076	0.544
MCTN [16]	32.7	75.9	76.4	0.991	0.613
MulT [22]	39.1	81.1	81.0	0.889	0.686
PMR [23]	40.6	82.4	82.1	–	–
MTAG [68]	38.9	82.3	82.1	0.866	0.722
MICA [24]	40.8	82.6	82.7	–	–
MFSA [12]	41.4	83.3	83.7	0.856	0.722
MCR (ours)	**42.6**†	**84.8**†	**84.6**†	**0.824**†	**0.735**†

Table 3
Comparison on the MOSEI benchmark. Best results are marked in **bold**.

Model	Acc_7 ↑	Acc_2 ↑	$F1$ ↑	MAE ↓	$Corr$ ↑
Word-aligned setting					
Graph-MFN [30]	45.0	76.9	77.0	0.710	0.540
RAVEN [15]	50.0	79.1	79.5	0.614	0.662
MCTN [16]	49.6	79.8	80.6	0.609	0.670
MFM [17]	49.8	78.5	78.8	0.633	0.657
RMFN [20]	50.7	79.1	79.5	0.619	0.672
MulT [22]	51.8	82.5	82.3	0.580	0.703
TCSP [18]	–	82.8	82.7	0.576	0.715
PMR [23]	52.5	83.3	82.6	–	–
FDMER [8]	53.8	83.9	83.8	0.568	0.736
MCR (ours)	**55.2**†	**85.1**†	**84.7**†	**0.542**†	**0.743**†
Unaligned setting					
RAVEN [15]	45.5	75.4	75.7	0.664	0.599
MCTN [16]	48.2	79.3	79.7	0.631	0.645
MulT [22]	50.7	81.6	81.6	0.591	0.694
PMR [23]	51.8	83.1	82.8	–	–
MICA [24]	52.4	83.7	83.3	–	–
MFSA [12]	53.2	83.8	83.6	0.574	0.724
MCR (ours)	**54.5**†	**84.7**†	**84.3**†	**0.554**†	**0.736**†

4.3.1. Word-aligned experiments

In this setting, the P2FA [69] is first utilized to align the acoustic and visual streams with the textual words. Then the multimodal fusion is conducted on the word-aligned time steps. The comparative results in the top half of Tables 2, 3, and 4 show that the proposed MCR significantly outperforms the previous methods in all evaluation metrics on the three benchmarks. It is worth noting that although our approach focuses on directly dealing with unaligned sequences, it still improves on most attributes by 6% to 9% over the original word-aligned methods [15–17,19,20].

4.3.2. Unaligned experiments

It is more challenging for the models to perform crossmodal fusion on unaligned multimodal sequences. The comparative results presented in the bottom half of Tables 2, 3, and 4 provide the following observations. Firstly, MCR is significantly superior to the models [15,16] that require CTC with a 10%–15% improvement on most attributes. In addition, our approach is more competitive and advantageous on all metrics for each benchmark

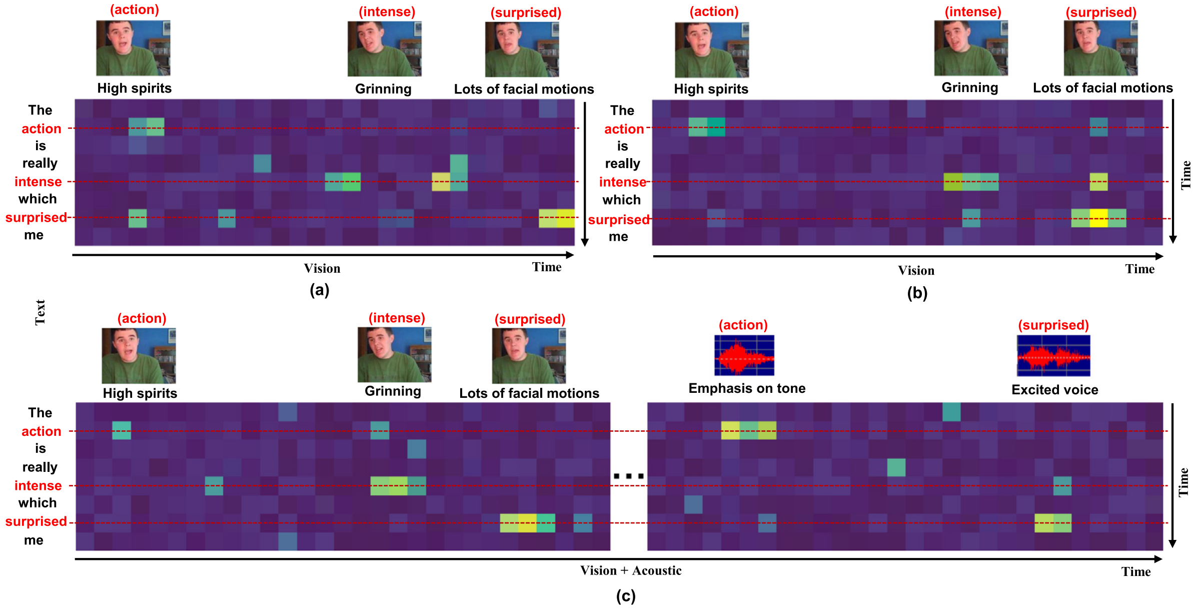

Fig. 3. Visualization of the crossmodal attention activation on the MOSI benchmark for different target modality reinforcement units. (a) The SOTA method MulT [22], (b) $CMU_{V,A\rightarrow L}$, and (c) $SCU_{[V,A]\rightarrow L}$ in our MCR.

Table 4
Comparison on the IEMOCAP benchmark. Best results are marked in **bold**.

Category	Happy		Sad		Angry		Neutral	
Model	*Acc* ↑	*F*1 ↑	*Acc* ↑	*F*1 ↑	*Acc* ↑	*F*1 ↑	*Acc* ↑	*F*1 ↑
Word-aligned setting								
Graph-MFN [30]	85.4	82.7	81.2	80.8	84.6	84.0	66.8	65.4
RMFN [20]	87.5	85.8	83.8	82.9	85.1	84.6	69.5	69.1
RAVEN [15]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
MCTN [16]	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
MFM [17]	90.2	85.8	88.4	86.1	87.5	86.7	72.1	68.1
MulT [22]	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7
PMR [23]	91.3	89.2	87.8	87.0	88.1	87.5	73.0	71.5
MCR (ours)	**92.6**†	**89.8**†	**88.6**†	**88.4**†	**89.4**†	**88.3**†	**74.1**†	**72.7**†
Unaligned setting								
RAVEN [15]	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
MCTN [16]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT [22]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
PMR [23]	86.4	83.3	78.5	75.3	75.0	71.3	63.7	60.9
MICA [24]	86.8	83.9	79.3	75.2	75.7	72.4	63.7	61.6
MCR (ours)	**87.4**†	**85.5**†	**81.4**†	**76.6**†	**76.1**†	**73.5**†	**66.3**†	**62.9**†

than the SOTA works [12,22–24] that directly deal with unaligned multimodal sequences. These observations suggest that our MCR can better perform multimodal fusion and interaction to learn valuable representations.

4.4. Visualization and analysis

4.4.1. Visualization of fine-grained interaction

To prove the superiority of $CMU_{s1,s2\rightarrow t}$ in modeling fine-grained crossmodal dependencies between elements, an example on the MOSI benchmark is chosen randomly. Figs. 3(a) & (b) show the crossmodal attention matrix activation for the last layer in the SOTA method MulT [22] (only MulT is open source) and the last layer of $CMU_{V,A\rightarrow L}$ in the proposed MCR, respectively. From Fig. 3(b), our module learns a reasonable correlation between the video frames and the spoken words. The emotion-related words (*e.g.*, "intense" and "surprised") successfully attend to the video frames which contain the corresponding facial expression changes (*e.g.*, "grinning" and "lots of facial motions"). Compared to MulT, our $CMU_{V,A\rightarrow L}$ encourages the model to focus on meaningful emotion signals in the interaction of independent modalities.

4.4.2. Visualization of mixed-grained interaction

In Fig. 3(c), we visualize the attention activation for the last layer of $SCU_{[V,A]\rightarrow L}$ from the same sample to show its crossmodal fusion performance. The proposed module attends more to the correlations between crossmodal elements that imply stronger emotion clues in mixed-grained interaction. For example, in the intersection of the emotionally-important word "action", stronger attention weights are given to audio clips that emphasize the tone rather than video frames. It is reasonable because the speaker's elevated intonation at this point is more reflective of the inspired emotion.

4.4.3. Reinforcement of crossmodal element correlations

To understand how the proposed components progressively reinforce the correlations between crossmodal elements, we visualize $CMU_{V,A\rightarrow L}$ and $SCU_{[V,A]\rightarrow L}$ on the first, second, and third layers, respectively. For clarity, Fig. 4 only shows the attention activation at the intersection of the most emotion-related word "bad" and the other elements from source modalities. We observe that both reinforcement units can gradually capture more meaningful crossmodal elements as the layers deepen. The conspicuous evidence is displayed on the third layer in which the word "bad" correlates well with the video frames suggesting negative emotion. These observations suggest that multi-grained information interactions lead to informative and effective multimodal representations.

4.4.4. Effectiveness of selective attention

We choose a sample each from the unaligned and word-aligned settings on the IEMOCAP benchmark to explain the selective attention weight α in mutual attention. From $CMU_{L,A\rightarrow V}$ in Fig. 5(a), the angry speaker says the emotion-related word "stupid", but her voice is neutral. There, the attention coefficient gives a lower weight to the audio modality, *i.e.*, $\alpha_A = 0.28$, and gives a higher weight to the language modality, *i.e.*, $\alpha_L = 0.72$. From $CMU_{L,V\rightarrow A}$ in Fig. 5(b), the happy speaker has a bright smile. However, there are no words in his description that clearly

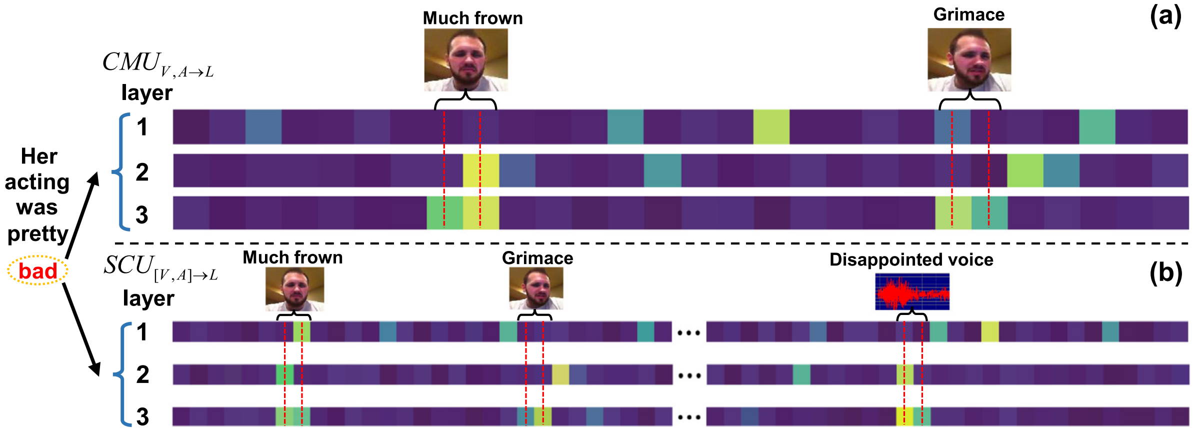

Fig. 4. Visualization of the local element attention activation on the MOSI benchmark for different layers of the proposed (a) $CMU_{V,A\to L}$ and (b) $SCU_{[V,A]\to L}$.

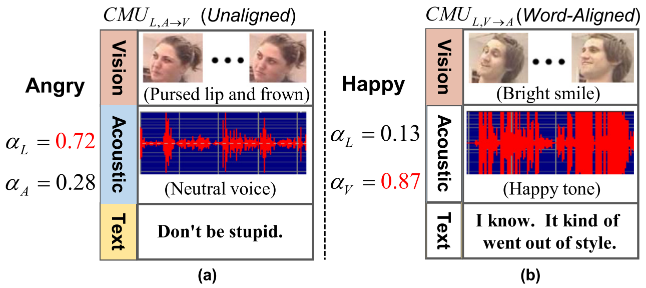

Fig. 5. Attention weights on the IEMOCAP benchmark. (a) Analysis of $CMU_{L,A\to V}$ in the unaligned setting. (b) Analysis of $CMU_{L,V\to A}$ in the word-aligned setting.

suggest emotion. The visual modality reasonably obtains a higher attention weight, *i.e.*, $\alpha_V = 0.87$, and the language modality obtains a lower attention weight, *i.e.*, $\alpha_L = 0.13$. The above observations demonstrate that the proposed mutual attention can assign reasonable weights to different modalities based on their importance. The adaptive weight pattern effectively highlights the stronger modalities while suppressing the weaker ones.

4.4.5. Comparison of convergence

We further explore the gain to the training process of the proposed strategies in different reinforcement units. These strategies include the Selective Memory Mechanism (SMM) in *SUM*, the mutual attention in $CMU_{s1,s2\to t}$, and the attention layer in $SCU_{[s1,s2]\to t}$. Fig. 6(a) shows the validation performance of our approach and the SOTA model MulT [22] in the unaligned setting on the MOSEI benchmark. Our MCR and its incomplete versions converge faster to obtain a lower mean average error than MulT. Meanwhile, the superiority of the full model benefits from the proposed strategies. The dynamic information filtering of the SMM is the most significant, as the convergence and performance of our model become worse when the SMM strategy is removed.

4.4.6. Confusion matrix

For the multi-class classification task on the IEMOCAP benchmark, Fig. 6(b) shows the confusion matrix to analyze the performance of each category. MCR effectively mitigates confusion between “happy” and “sad”. Another finding is that some data samples are confused between “happy” and “angry” because people tend to exaggerate emotion cues in both cases. The phenomenon is consistent with previous work [70].

Table 5
Ablation study results on the MOSI and MOSEI benchmarks in the unaligned setting. SMM refers to the Selective Memory Mechanism in *SUM*. FFL means Feed-Forward Layer.

Model	MOSI			MOSEI		
	Acc_2 ↑	$F1$ ↑	MAE ↓	Acc_2 ↑	$F1$ ↑	MAE ↓
MCR (Full model)	**84.8**	**84.6**	**0.824**	**84.7**	**84.3**	**0.554**
Importance of modality						
Language Only	80.6	79.8	0.855	81.4	80.8	0.581
Video Only	77.8	76.1	0.897	78.6	77.2	0.595
Audio Only	76.2	75.8	0.914	78.2	76.9	0.613
Importance of crossmodal interaction						
$[V, A \to L]$ Only	83.1	82.7	0.833	83.2	82.8	0.568
$[L, A \to V]$ Only	81.6	81.2	0.841	81.6	81.1	0.578
$[L, V \to A]$ Only	81.2	80.4	0.850	81.1	80.4	0.585
Effectiveness of source modality reinforcement						
Low-level Feature	82.5	82.3	0.836	82.4	81.8	0.572
High-level Feature (FFL)	81.8	81.6	0.840	81.7	81.3	0.576
SUM w/o SMM	83.2	82.9	0.832	83.5	83.0	0.564
Effectiveness of target modality reinforcement						
w/o $CMU_{s1,s2\to t}$	81.3	81.0	0.845	82.6	82.3	0.570
w/o Mutual Attention	84.1	83.7	0.827	83.8	83.4	0.562
w/o $SCU_{[s1,s2]\to t}$	82.2	81.8	0.838	81.3	80.6	0.582
w/o Attention Layer	83.5	83.2	0.832	84.2	83.7	0.559
Rationality of structure						
w/ Serial Structure	82.1	81.5	0.839	81.8	81.6	0.574

4.5. Ablation study

In Table 5, we perform thorough ablation experiments in the unaligned setting to verify the necessity of all proposed components on the MOSI and MOSEI benchmarks. All results are obtained from the average of five identical experiments to avoid randomness.

4.5.1. Importance of modality

Exploring the performance of a single modality is essential for a multimodal system. To this end, we first evaluate the unimodal MCR’s performance. In this setup, the crossmodal attention among the modalities of $CMU_{s1,s2\to t}$ and $SCU_{[s1,s2]\to t}$ is equivalent to the self-attention within the modality. The significant performance drops suggest that the isolated modality provides limited improvement and emotional clues. Another finding is that the unimodal MCR based on language modality outperforms the other two on all metrics by a large margin. For the Acc_2 metric, the model improves from 76.2% to 80.6% when comparing the audio-only to language-only MCR. This fact aligns with the observations

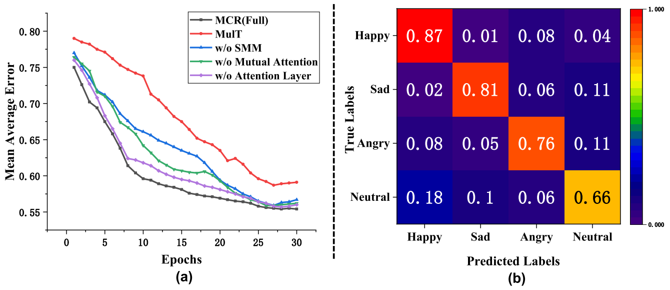

Fig. 6. (a) The validation set convergence and mean average error of our approach compared to the SOTA method MulT on the unaligned MOSEI task. (b) Confusion matrix for emotion classification on the unaligned IEMOCAP task.

in prior work [16], where the authors found that a good language model can achieve adequate performance.

4.5.2. Importance of crossmodal interaction

For crossmodal MCR, we keep either of the crossmodal combinations separately (*i.e.*, $[V, A \rightarrow L]$, $[L, A \rightarrow V]$, and $[L, V \rightarrow A]$). The decreased results on all metrics suggest that it is beneficial to consider these three crossmodal interactions concurrently. Moreover, the crossmodal combination where language is the target modality works best. One reasonable explanation is that acoustic and visual features contain more noisy and redundant information than textual features [62], limiting the model's performance.

4.5.3. Effectiveness of source modality reinforcement

We first provide two replacements for the low-level and high-level features, respectively. Specifically, the low-level version leverages the features from the initial multimodal sequences. The high-level version only reinforces the source modalities by stacking the feed-forward layer. The worst performance is observed in the high-level version. The possible reason is that the updates of the source modalities are independent and limited, which produces ineffective representations to confuse the reinforced target modalities. This fact aligns with the observations in previous works [22,23]. In contrast, the proposed storage neuron-inspired *SUM* can receive explicit supervision of the target modalities to update more meaningful high-level features. Furthermore, when removing the SMM from *SUM*, the input features are merged by the element-wise sum. The poor results on all metrics reveal that information filtering is essential.

4.5.4. Effectiveness of target modality reinforcement

To explore the importance of target modality reinforcement at different granularities, we respectively remove $CMU_{s1,s2\rightarrow t}$ and $SCU_{[s1,s2]\rightarrow t}$ to perform experiments using only the partially reinforced features. The performance degradation suggests that it is beneficial to consider fine-grained and mixed-grained interactions in crossmodal fusion. Furthermore, we find that both the mutual attention and the attention layer provide indispensable contributions to $CMU_{s1,s2\rightarrow t}$ and $SCU_{[s1,s2]\rightarrow t}$, respectively.

4.5.5. Rationality of structure

Evaluating the tailored module combination paradigm in Fig. 1(b) plays an important role in the model structure design. To this end, the default parallel structure of the two target reinforcement units is replaced with an alternating serial structure to explore the impact on model performance. The decreased results from the bottom of Table 5 reveal the effectiveness and superiority of the proposed MCR. The strengths of the current structure may benefit from mimicking the parallel signal-processing pattern of the sensory neurons in the cerebral cortex [28].

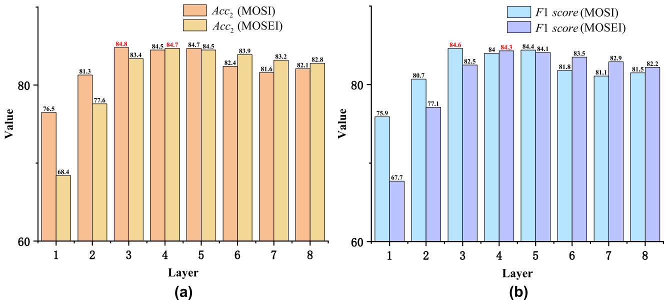

Fig. 7. Sensitivity analysis on the unaligned MOSI and MOSEI benchmarks. We show the effect of the number of modality reinforcement layers on performance using (a) Acc_2 and (b) $F1$ score metrics. The red values represent the best performance.

4.5.6. Sensitivity of modality reinforcement layers

Fig. 7 shows the effect of varying the number of layers in $CMU_{s1,s2\rightarrow t}$ and $SCU_{[s1,s2]\rightarrow t}$ on performance. We observe that the performance gradually improves as the number of layers increases and then tends to saturate. The proposed approach achieves optimal gain when the number of layers on the MOSI and MOSEI benchmarks is 3 and 4, respectively. Compared to the previous SOTA models [8,12,22–24] with extensive parameters and module stacking, MCR achieves superior performance with fewer parameters (*i.e.*, a smaller number of layers). Another finding is that too many modules stacked can cause gain drops. One potential reason is that the deeply dense structure increases the complexity and thus limits the model's performance.

5. Conclusion

Researching multimodal models driven by the human perceptual system is essential to bridging the perceptual gap between humans and machines. To this end, we present a target and source modality co-reinforcement approach to achieve multimodal fusion from unaligned multimodal sequences for emotion understanding. Our target modality reinforcement units thoroughly explore the inherent correlations between the elements of different modalities via fine-grained and mixed-grained interactions. The source modality update module progressively provides reliable high-level representations for crossmodal fusion, benefiting from the proposed selective memory mechanism. Numerous experiments over different benchmarks demonstrate the superiority of MCR. The effective components of our model can serve diverse applications, *e.g.*, complex industrial scenarios that require multi-sensor signal fusion. The future direction is to explore how to employ MCR on other multimodal understanding tasks.

CRediT authorship contribution statement

Dingkang Yang: Conceptualization, Methodology, Writing – original draft. **Yang Liu:** Writing – review & editing. **Can Huang:** Formal analysis. **Mingcheng Li:** Resources. **Xiao Zhao:** Validation. **Yuzheng Wang:** Investigation. **Kun Yang:** Software. **Yan Wang:** Visualization. **Peng Zhai:** Supervision. **Lihua Zhang:** Project administration, Funding acquisition.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

We thank Shunli Wang, Liuzhen Su, Zhaoyu Chen, Haopeng Kuang, and the anonymous reviewers for providing constructive discussions and feedbacks. This work is supported in part by the National Key R&D Program of China (2021ZD0113502, 2021ZD0113503), in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103), and in part by the China Postdoctoral Science Foundation under Grant (BX20220071, 2022M720769).

References

[1] W. Chen, P. Lin, W. Zhang, J. Du, Z. He, Hierarchical Interactive Network for joint aspect extraction and sentiment classification, Knowl.-Based Syst. 256 (2022) 109825.

[2] X. Wang, M. Fan, M. Kong, Z. Pei, Sentiment Lexical Strength Enhanced Self-supervised Attention Learning for sentiment analysis, Knowl.-Based Syst. 252 (2022) 109335.

[3] G. Wen, H. Liao, H. Li, P. Wen, T. Zhang, S. Gao, B. Wang, Self-labeling with feature transfer for speech emotion recognition, Knowl.-Based Syst. 254 (2022) 109589.

[4] Q. Lu, X. Sun, R. Sutcliffe, Y. Xing, H. Zhang, Sentiment interaction and multi-graph perception with graph convolutional networks for aspect-based sentiment analysis, Knowl.-Based Syst. 256 (2022) 109840.

[5] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Comput. Vis. Image Underst. 174 (2018) 33–42, http://dx.doi.org/10.1016/j.cviu.2018.06.005.

[6] W. Aljedaani, F. Rustam, M.W. Mkaouer, A. Ghallab, V. Rupapara, P.B. Washington, E. Lee, I. Ashraf, Sentiment analysis on twitter data integrating textblob and deep learning models: the case of us airline industry, Knowl.-Based Syst. 255 (2022) 109780.

[7] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, S. Levinson, Audio-visual affect recognition through multi-stream fused HMM for HCI, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2005, pp. 967–972.

[8] D. Yang, S. Huang, H. Kuang, Y. Du, L. Zhang, Disentangled representation learning for multimodal emotion recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1642–1651.

[9] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, L. Zhang, Emotion recognition for multiple context awareness, in: Proceedings of the European Conference on Computer Vision, Vol. 13697, Springer, 2022, pp. 144–162.

[10] Y. Du, D. Yang, P. Zhai, M. Li, L. Zhang, Learning associative representation for facial expression recognition, in: IEEE International Conference on Image Processing, 2021, pp. 889–893.

[11] Y.-H.H. Tsai, M.Q. Ma, M. Yang, R. Salakhutdinov, L.-P. Morency, Multimodal routing: Improving local and global interpretability of multimodal language analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, NIH Public Access, 2020, p. 1823.

[12] D. Yang, H. Kuang, S. Huang, L. Zhang, Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1708–1717.

[13] D. Yang, S. Huang, Y. Liu, L. Zhang, Contextual and cross-modal interaction for multi-modal speech emotion recognition, IEEE Signal Process. Lett. 29 (2022) 2093–2097.

[14] Z. Zhang, J.M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al., Multimodal spontaneous emotion corpus for human behavior analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

[15] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7216–7223.

[16] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6892–6899.

[17] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, in: International Conference on Representation Learning, 2018.

[18] Y. Wu, Z. Lin, Y. Zhao, B. Qin, L.-N. Zhu, A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics, 2021, pp. 4730–4738.

[19] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88, http://dx.doi.org/10.1109/MIS.2016.94.

[20] P.P. Liang, Z. Liu, A. Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 150–161.

[21] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 369–376.

[22] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting. 2019, NIH Public Access, 2019, p. 6558.

[23] F. Lv, X. Chen, Y. Huang, L. Duan, G. Lin, Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2554–2562.

[24] T. Liang, G. Lin, L. Feng, Y. Zhang, F. Lv, Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8148–8156.

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

[26] J. Zhu, X. Zhang, R. Wang, M. Wang, P. Chen, L. Cheng, Z. Wu, Y. Wang, Q. Liu, M. Liu, A heterogeneously integrated spiking neuron array for multimode-fused perception and object classification, Adv. Mater. (2022) 2200481.

[27] H. Yang, X. Li, Y. Wu, S. Li, S. Lu, J.S. Duncan, J.C. Gee, S. Gu, Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 799–807.

[28] K. Zilles, K. Amunts, Receptor mapping: architecture of the human cerebral cortex, Cur. Opi. Neu. 22 (4) (2009) 331–339.

[29] L. Fernandino, J.R. Binder, R.H. Desai, S.L. Pendl, C.J. Humphries, W.L. Gross, L.L. Conant, M.S. Seidenberg, Concept representation reflects multimodal abstraction: A framework for embodied semantics, Cereb. Cortex 26 (5) (2016) 2018–2034.

[30] A. Zadeh, P. Pu, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2236–2246.

[31] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.

[32] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, D. Manocha, Step: Spatial temporal graph convolutional networks for emotion perception from gaits, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 1342–1350.

[33] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Sign. Proces. 11 (8) (2017) 1301–1309, http://dx.doi.org/10.1109/JSTSP.2017.2764438.

[34] B. Xie, M. Sidulova, C.H. Park, Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion, Sensors 21 (14) (2021) 4913, http://dx.doi.org/10.3390/s21144913.

[35] Y. Liu, J. Liu, M. Zhao, D. Yang, X. Zhu, L. Song, Learning appearance-motion normality for video anomaly detection, in: 2022 IEEE International Conference on Multimedia and Expo, 2022, pp. 1–6.

[36] Y. Liu, J. Liu, X. Zhu, D. Wei, X. Huang, L. Song, Learning task-specific representation for video anomaly detection with spatial-temporal attention, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 2190–2194.

[37] Y. Huang, H. Wen, L. Qing, R. Jin, L. Xiao, Emotion recognition based on body and context fusion in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3609–3617.

[38] W. Duan, L. Zhang, J. Colman, G. Gulli, X. Ye, Multi-modal brain segmentation using hyper-fused convolutional neural network, in: Lect. Notes Comput. Sci., Springer, 2021, pp. 82–91.

[39] Z. Chen, B. Li, J. Xu, S. Wu, S. Ding, W. Zhang, Towards practical certifiable patch defense with vision transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15148–15158.

[40] H. Huang, Y. Wang, Z. Chen, Y. Zhang, Y. Li, Z. Tang, W. Chu, J. Chen, W. Lin, K.-K. Ma, Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 989–997.

[41] P. Zhai, J. Luo, Z. Dong, L. Zhang, S. Wang, D. Yang, Robust adversarial reinforcement learning with dissipation inequation constraint, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022 pp. 5431–5439.

[42] P. Zhai, T. Hou, X. Ji, Z. Dong, L. Zhang, Robust adaptive ensemble adversary reinforcement learning, IEEE Robot. Autom. (2022).

[43] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning, 2011.

[44] A. Lazaridou, N.T. Pham, M. Baroni, Combining language and vision with a multimodal skip-gram model, 2015, arXiv preprint arXiv:1501.02598.

[45] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A.N. Vembu, R. Prasad, Emotion recognition using acoustic and lexical features, in: Proceedings of Interspeech, 2012, 2012, pp. 366–369.

[46] H. Ranganathan, S. Chakraborty, S. Panchanathan, Multimodal emotion recognition using deep learning architectures, in: 2016 IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–9.

[47] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, C. Fookes, Deep spatio-temporal features for multimodal emotion recognition, in: 2016 IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 1215–1223.

[48] D. Yang, H. Shuai, S. Wang, P. Zhai, Y. Li, L. Zhang, EE-GAN: facial expression recognition method based on generative adversarial network and network integration, J. Comput. Appl. 42 (3) (2022) 750–756.

[49] Y.R. Pandeya, J. Lee, Deep learning-based late fusion of multimodal information for emotion classification of music video, Multimedia Tools Appl. 80 (2) (2021) 2887–2905.

[50] H. Devamanyu, Z. Roger, P. Soujanya, MISA: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[51] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, 2017, arXiv preprint arXiv:1707.07250.

[52] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, 2018, arXiv preprint arXiv:1806.00064.

[53] L.R. Medsker, L. Jain, Recurrent neural networks, Des. Appl. 5 (2001) 64–67.

[54] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[55] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. 27 (2014) 3104–3112.

[56] Z. Chen, B. Li, S. Wu, J. Xu, S. Ding, W. Zhang, Shape matters: deformable patch attack, in: Proceedings of the European Conference on Computer Vision, Springer, 2022, pp. 529–548.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[58] X. Chen, Y. Wu, Z. Wang, S. Liu, J. Li, Developing real-time streaming transformer transducer for speech recognition on large-scale dataset, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 5904–5908.

[59] L. Dong, S. Xu, B. Xu, Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5884–5888.

[60] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C.A. Hunter, C. Bekas, A.A. Lee, Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction, ACS Cent. Sci. 5 (9) (2019) 1572–1583.

[61] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.

[62] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 163–171.

[63] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 873–883.

[64] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/D14-1162.

[65] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP—A collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2014, pp. 960–964, http://dx.doi.org/10.1109/ICASSP.2014.6853739.

[66] iMotions, Facial Expression Analysis, 2017.

[67] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[68] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, L.-P. Morency, Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2020.

[69] J. Yuan, M. Liberman, et al., Speaker identification on the SCOTUS corpus, J. Acoust. Soc. Am. 123 (5) (2008) 3878.

[70] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 1359–1367.