MIA-Net: Multi-modal Interactive Attention Network For Multi-modal Affective Analysis

Shuzhen Li, Tong Zhang*, *Member, IEEE,* Bianna Chen, C. L. Philip Chen, *Fellow, IEEE*

Abstract—When a multi-modal affective analysis model generalizes from a bimodal task to a trimodal or multi-modal task, it is usually transformed into a hierarchical fusion model based on every two pairwise modalities, similar to a binary tree structure. This easily leads to large growth in model parameters and computation as the number of modalities increases, which limits the model's generalization. Moreover, many multi-modal fusion methods ignore that different modalities contribute differently to affective analysis. To tackle these challenges, this paper proposes a general multi-modal fusion model that supports trimodal or multi-modal affective analysis tasks, called Multi-modal Interactive Attention Network (MIA-Net). Instead of treating different modalities equally, MIA-Net takes the modality that contributes the most to emotion as the main modality and the others as auxiliary modalities. MIA-Net introduces multi-modal interactive attention modules to adaptively select the important information of each auxiliary modality one by one to improve the main-modal representation. Moreover, MIA-Net enables quick generalization to trimodal or multi-modal tasks through stacking multiple MIA modules, which maintains efficient training and only requires linear computation and stable parameter counts. Experimental results of the transfer, generalization, and efficiency experiments on the widely-used datasets demonstrate the effectiveness and generalization of the proposed method.

Index Terms—Multi-modal affective analysis, multi-modal fusion, multi-modal interactive attention.

1 Introduction

With the development of the Internet and intelligent terminal equipment such as smartphones, more and more people spend much time in interacting with machines. Human-computer interaction has become an essential part of our lives. To achieve a more friendly and natural human-machine interaction, the machine should be able to understand the user's emotions. Affective analysis mainly includes sentiment analysis and emotion recognition [1]. Sentiment analysis aims to determine the attitude of a speaker or a writer towards some topic or some documents [1]. It is usually considered a polarity classification or the strength detection of polarity. The polarity can be expressed as positive, negative, or neutral. Emotion recognition performs fine-grained affective recognition according to emotional labels [2]. These emotional labels include anger, disgust, fear, happiness, neutral, sadness, surprise, excitement, etc. Affective analysis plays a crucial role in natural human-computer interaction and has a wide range of practical applications. For instance, doctors utilize affective analysis technology to research and treat mental illnesses, such as depression [3], autism [4], [5], stress [6], etc. Sentiment analysis can be used to monitor and predict the drivers' fatigue state [7]. Effective emotion recognition can bring a more comfortable interaction experience to users in the service industry, such as customer service, online education, etc. Therefore, affective analysis has attracted much attention in recent years.

Much effort has focused on analyzing emotions from a single modality such as text [8], [9], [10], video [11], [12], [13], and audio [14], [15], [16]. However, the robustness of unimodal affective analysis systems is weak due to the incomplete information. Furthermore, some representations can be ambiguous under unimodal affective analysis systems. For example, when someone in a dilemma is forced to laugh, the real emotional information may hide in voice or linguistic content. Also, it is difficult to portray the true emotion of a sad or angry person from neutral text descriptions, such as " I'm OK ". Therefore, it is necessary to conduct multi-modal affective analysis to solve the ambiguity and retain complimentary affective information. Thus, this paper focuses on multi-modal affective analysis.

The main challenge in multi-modal affective analysis is how to adjust bimodal model to quick generalization to trimodal or multi-modal data tasks. Many researchers attempted to address this problem via hierarchical fusion. In hierarchical fusion, bimodal interactions are generated based on unimodal information, and trimodal interactions

- *Shuzhen Li is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China.*
- *Tong Zhang is with the Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and is with the Pazhou Lab, Guangzhou 510335, China.*
  E-mail: tony@scut.edu.cn
- *Bianna Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China.*
- *C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China.*

*This work was funded in part by the National Key Research and Development Program of China under number 2019YFA0706200, in part by the National Natural Science Foundation of China grant under number 62222603, 62076102, U1813203, and U1801262, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under number 2020B1515020041, in part by the Science and Technology Major Project of Guangzhou under number 202007030006, in part by the Science and Technology Program of Guangzhou under number 202002030250, in part by The Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08X214). (*Corresponding author:Tong Zhang)*

are obtained based on bimodal information [17]. A contextual inter-modal (CIM) attention mechanism [18] was presented to learn the relations among neighboring utterances and every two modalities, while it fused all bimodal features via concatenation. A unified multi-modal emotion set generation network [18] employed a cross-modal transformer encoder to learn cross-modal features between every two modalities and concatenated all cross-modal features as a hybrid representation. A novel multi-interactive memory network model [19] was proposed to learn each bimodal representation by attention mechanism and adopted the concatenation as the late fusion method. A multi-modal transformer (MulT) [20] merged multi-modal timeseries via multiple directional pairwise cross-modal transformers, and modeled all pairwise fusion features with such cross-modal transformers and concatenation. These hierarchical fusion methods learned the bimodal features from every two pairwise modalities, and finally fused all bimodal features by simple concatenation. As the interactions between every two pairwise modalities are different, simple concatenation fails to mine the inter-bimodal relationships [17]. Moreover, the hierarchical fusion model based on every two pairwise modalities easily leads to large growth in model parameters and computation as the number of modalities increases, which limits the model's generalization. Therefore, this paper designs an effective fusion strategy that enables quick generalization to the trimodal or multi-modal tasks under limited computation and parameters.

Different organs' perceptions provide different degrees of reference for the human brain when human analyzes emotions. It is rational that not all modalities play equal roles in affective analysis [21]. However, most previous works [22], [23], [24] treated three modalities equally, which might weaken the performance of multi-modal systems [25]. As highly semantic information, text descriptions can express more explicit emotion than other modalities and contribute more to affective analysis [26]. Moreover, some works [27], [28], [29] indicated that language modality is far more informative than the visual and acoustic modalities, which should play a dominant role when learning the joint representation. Therefore, this paper considers the modality that contributes the most to emotion as the main modality and the others as auxiliary modalities, and designs a multi-modal fusion method based on the main and auxiliary modalities.

To address the above challenges, this paper proposes a general multi-modal affective analysis model called multi-modal interactive attention network (MIA-Net). In MIA-Net, the modality that contributes the most to emotion is treated as the main modality and the other modalities as the auxiliary modalities. MIA-Net mainly consists of $N-1$ multi-modal interactive attention (MIA) modules for affective analysis tasks with N modalities. Each multi-modal interactive attention module explores the important relationships of each pairwise modalities and encodes these interactions as attention weights to improve main-modal representation. MIA-Net can maintain efficient training and require linear computation and stable parameter counts when stacking multiple MIA modules to add more modalities. This enables quick generalization to trimodal or multi-modal tasks. Finally, the proposed approach achieves state-of-the-art performance, transfer learning performance, and generalization performance across several widely used multi-modal affective analysis datasets, demonstrating the effectiveness and generalization of MIA-Net.

The main contributions are summarized as follows:

- This paper proposes a general multi-modal fusion method for affective analysis, named multi-modal interactive attention network. It enables quick generalization to trimodal or multi-modal tasks under linear computation and stable parameter counts.
- This paper presents a general linear kernel as the similarity computation in multi-modal interactive attention, which adaptively learns the interaction relationship of different pairwise modalities along with the training of the network and remains the multi-modal feature norm unchanged.

The rest of this paper is organized as follows. Section 2 provides a brief literature review of prior sentiment analysis and multi-modal fusion methods. Section 3 describes the proposed approach in detail. Next, the experimental results and analysis are elaborated in Section 4 and Section 5. Finally, Section 6 makes conclusions and describes the future work.

2 Related Work

This section focuses on the research of current affective analysis methods and multi-modal fusion methods.

2.1 Affective Analysis

Affective analysis mainly includes sentiment analysis and emotion recognition [1]. Sentiment analysis aims to determine the attitude of a speaker or a writer towards some topic or some documents [1]. It is usually considered a polarity classification or the strength detection of polarity, such as positive, negative, or neutral. Emotion recognition performs fine-grained affective recognition according to emotional labels [2]. These emotional labels include anger, disgust, fear, happiness, neutral, sadness, surprise, excitement, etc. Most of the previous works conducted affective analysis using unimodal features [2]. For instance, the pre-trained end-to-end speech recognition model extracted acoustic features for speech sentiment analysis [30]. A distantly supervised lifelong sentiment learning framework was proposed to detect the sentiments of continuously updated texts with dynamic topics in social media [31]. A deep learning approach employed bidirectional long-short term memory (Bi-LSTM) and recurrent neural networks (RNNs) to detect human facial sentiment [32]. Since multi-modal learning exploited the complementary information of different modalities to improve model performance, multi-modal learning was introduced into affective analysis and attracted much attention. A novel locally-confined modality fusion network (LMFN) was proposed for multi-modal sentiment analysis, which decomposed the feature vector of each modality into multiple feature segments and learned local interactions [2]. Inspired by cognitive neuroscience, a hierarchical attention-LSTM model based on the cognitive brain limbic system was presented for multi-modal sentiment analysis [33]. More

recently, some multi-modal sentiment analysis and multi-modal emotion recognition datasets have been published and significantly promoted the development of multi-modal affective analysis, such as CMU-MOSI [34], CMU-MOSEI [35] and MELD [36].

2.2 Multi-modal Fusion

Multi-modal fusion means integrating information from multiple modalities to predict discrete classes by classification or continuous values by regression [37]. It is the original topic in multi-modal machine learning. Early research on multi-modal fusion was divided into signal-level, feature-level, and decision-level fusion [38]. As one of the most popular topics in multi-modal fusion, the feature-level fusion method mixes together the features outputted by different signal processors. Early works [39], [40], [41] conducted feature-level fusion most by simple concatenation. However, simple concatenation was unsuitable for multi-modal heterogeneous features and did not allow efficient intra-modality dynamics [42], [43].

Recent works on feature-level fusion can be divided into the following categories.

Tensor-based fusion method aims to perform mathematical operations on each modal tensor to learn inter-modality dynamics. Tensor fusion network (TFN) [42] performed the outer product between multi-modal tensors to aggregate multi-modal interactions. Low-rank multi-modal fusion (LMF) model [23] conducted multi-modal fusion by low-rank tensors to improve the efficiency of TFN model. A multi-modal factorization model [44] was proposed to learn multi-modal embedding via factorization methods. However, the joint representations from the tensor-based fusion methods are much sparse [45], and tensor-based fusion methods did not explicitly outstand the dominant role of the main modality [25].

Modality-invariant representation learning method aims to utilize the specific loss functions to constrain multi-modal features to learn the shared representations. Ma *et al.* [46] employed a correlation loss and a reconstruction loss to preserve the shared and private information, respectively. In order to maintain task-related information through multi-modal fusion, a MultiModal InfoMax (MMIM) framework [47] was proposed to hierarchically maximize the mutual information in unimodal input pairs and between multi-modal fusion results and unimodal input. A multi-modal information bottleneck (MIB) [48] was presented to learn the minimal sufficient representation by maximizing the mutual information between the representation and the target and constraining the mutual information between the representation and the input data. A novel multi-modal framework MISA [49] learned modality-invariant and modality-specific features by similarity loss and difference loss, respectively. Although modality-invariant representation learning methods have significantly contributed to multi-modal fusion, they suffered from weak interpretability.

Attention-based fusion method aims to exploit attention mechanisms to learn multi-modal interactions. Some recent works adopted soft attention [50] to dynamically adjust the weights of multi-modal features, such as CM-BERT [51], DialogueRNN [52], A-DMN [53], multi-modal routing model [54]. Since multi-head attention (MHA) [55] allowed the model to jointly attend to information from different modal representations, the bi-bimodal fusion network (BBFN) [56] exploited MHA to extract the multi-modal information. A self-supervised learning (SSL) model [57] for multi-modal sentiment analysis fused acoustic and textual features via MHA and concatenation technique. Transformer [55], consisting of multiple MHA, succeeded in drawing global dependencies between two inputs in machine translation task, so it was introduced into multi-modal learning and attracted much attention. A conversational transformer network (CTNet) [58] employed the Transformer to model intra-modal and cross-modal interactions among multi-modal features. A unified multi-modal emotion set generation network [59] built various cross-modal transformer encoders to capture the cross-modal interactions among textual, visual, and acoustic modalities. However, the Transformer-based fusion method used the decoder component of the standard Transformer to improve performance, which may lead to some redundancy [60]. In addition, other attention variants have been applied to multi-modal fusion. A temporal convolutional multi-modal LSTM (TCM-LSTM) [25] introduced the sparse attention to determine whether the acoustic and visual enhancement should be conducted. A contextual inter-modal (CIM) attention mechanism [18] was presented to learn the degree of association among neighboring utterances and the input modalities.

3 Methodology

This section presents a multi-modal interactive attention network for multi-modal affective analysis. Fig. 1 depicts the framework of the proposed method. The framework mainly consists of feature extraction, multi-modal interactive attention network, and a regressor or classifier. The details are described below.

3.1 Feature Extraction

Many modalities can express emotion in affective analysis tasks. This paper takes the modalities from the CMU-MOSI, CMU-MOSEI, and MELD datasets as examples, including text, audio, and video modalities. This subsection extracts textual, acoustic, and visual features from the text, audio, and video data through a series of data pre-processing and feature extraction.

3.1.1 Text data

Since the text description is highly semantic information, it is challenging to extract effective textual sentiment features via simple manual feature extraction methods, such as splitting sentences, fixed words, word segmentation, markers, etc [61]. So this paper adopts the automatic textual feature extraction method based on deep learning models. Firstly, each text description is tokenized into a sequence of tokens by GPT-2 tokenizer [62]. Then, the tokens are fed into a pre-trained RoBERTa model to extract effective textual features automatically. As one of the BERT [63] variants, the RoBERTa model [64] learned deep bidirectional language representations by jointly conditioning on both left and right

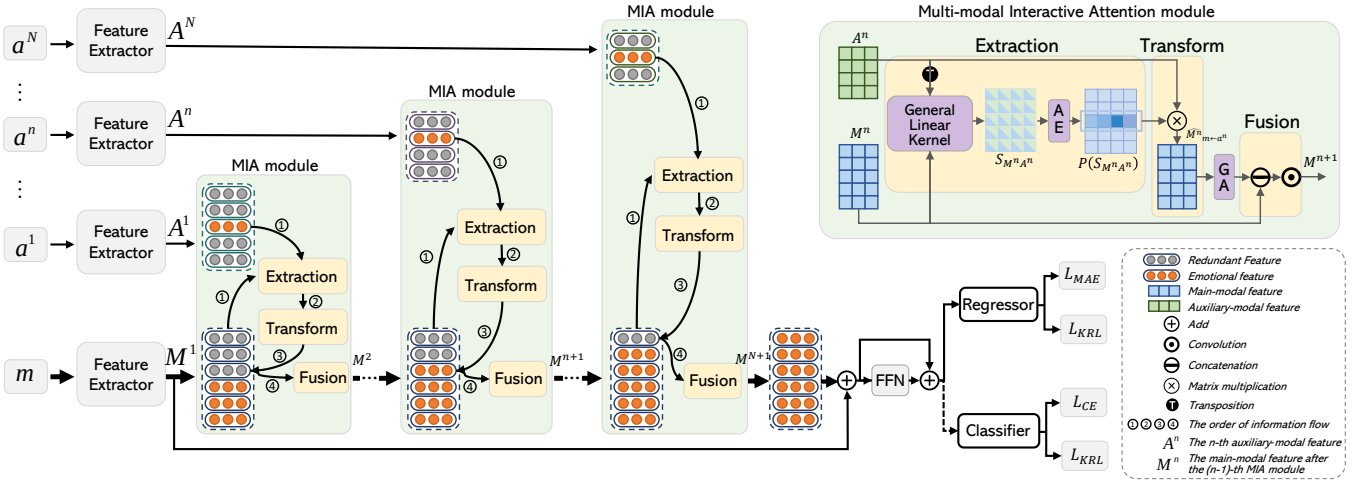

Fig. 1. The framework of the proposed MIA-Net for multi-modal affective analysis. In a multi-modal affective analysis task with $N+1$ modalities, the proposed framework mainly consists of unimodal feature extractors, N multi-modal interactive attention modules, and a regressor or classifier. Each MIA module is mainly composed of Extraction, Transformation, and Fusion submodules, and aims to select important features from auxiliary modalities to improve the main-modal representation.

context in all layers and using dynamic masking. The pre-trained RoBERTa model consists of 24 transformer layers and an output embedding layer with 1024 neurons. Therefore, this paper extracts textual features with a dimension of 1024 from each text description.

3.1.2 Audio data

In order to maintain the integrity of acoustic information, this paper employs the variable-length feature extraction method to extract acoustic features from variable-length audio. Firstly, this work uses a pre-trained model VQ-Wav2Vec to represent each variable-length audio as a sequence of discrete acoustic representations. VQ-Wav2Vec [65] was a model that quantized variable-length audio data, making it amenable to algorithms requiring discrete data. The pre-trained model VQ-Wav2Vec is trained on the Librispeech-960 dataset [66]. Secondly, this work adopts a pre-trained RoBERTa model to extract acoustic features from the discrete acoustic representations. The pre-trained RoBERTa model comprises 12 transformer layers and an output embedding layer with 768 neurons, and it is trained on the discretized Librispeech-960 dataset with the mask token prediction task. Therefore, this paper extracts acoustic features with a dimension of 768 from each variable-length audio data.

3.1.3 Video data

In this work, the pre-processing of video data is described as follows: (1) Facial region detection and segmentation. In each video frame, sentiment information is mainly distributed on human faces rather than the background. So this work employs the RetinaFace model to detect and segment the facial region in each frame. RetinaFace [67] was a robust single-stage face detector that performed pixel-wise face localization on various scales of faces. (2) Keyframes extraction. The number of frames varies greatly from video to video. This work sets the number of keyframes per video as the average of video frames across the whole dataset. Then, this work extracts video frames from each video at equal intervals according to the number of keyframes. (3) Visual feature extraction. The keyframes are fed into a pre-trained Fabnet model to extract visual features. The Fabnet model [68] learned facial attributes and embedded multiple frames into a common low-dimensional space with dimension 256. Therefore, this paper extracts visual features with a dimension of 256 from each video.

3.2 Multi-modal Interactive Attention Network

In an affective analysis task with N modalities, the multi-modal interactive attention network takes one modality as the main modality and other modalities as the auxiliary modalities, and it stacks $N-1$ multi-modal interactive attention modules to make important information flow from the auxiliary modalities to the main modality. The features from the last multi-modal interactive attention module are then fed into a dual residual structure to get more discriminative features. The details are as follows.

3.2.1 Multi-modal Interactive Attention Module

The proposed multi-modal interactive attention module has two properties: 1) It can adaptively capture important relations of different pairwise modalities; 2) It can maintain efficient training and less computational growth when stacking N MIA modules to expand N modalities. Thus, this work designs a multi-modal interactive attention and a kernel regularization loss in each multi-modal interactive attention module. This paper denotes the main modality and the n-th auxiliary modality as m and a^n, respectively.

Multi-modal Interactive attention. Interactive attention [69] is an attention mechanism that can calculate a comprehensive attention weight via two input features and denote the correlation between the two input features. Inspired by interactive attention, this work proposes multi-modal interactive attention for multi-modal data. It can adaptively capture important relations of different pairwise modalities and encode the interactive relations as interactive attention weights for improving the main-modal feature representation.

The main-modal features of the n-th multi-modal interactive attention module are denoted as $\boldsymbol{M}^n \in \mathbb{R}^{c\times d_m}$, and the n-th auxiliary-modal features are denoted as $\boldsymbol{A}^n \in \mathbb{R}^{c\times d_{a^n}}$. d_m and d_{a^n} are the dimensions of main-modal features and auxiliary-modal features, respectively. Each column $\boldsymbol{M}_j^n \in \mathbb{R}^c, j \in \{1, ..., d_m\}$ represents the j-th main-modal feature vector of $\boldsymbol{M}^n$. Each column $\boldsymbol{A}_i^n \in \mathbb{R}^c, i \in \{1, ..., d_{a^n}\}$ represents the i-th auxiliary-modal feature vector of $\boldsymbol{A}^n$.

Extraction. Given the main-modal features $\boldsymbol{M}^n$ and the n-th auxiliary-modal features $\boldsymbol{A}^n$, the interactive relations between all main-modal and auxiliary-modal feature vector pairs are first explored. Since the similarity computation method measures the importance of two inputs, it is feasible to introduce the similarity computation method to multi-modal interactive attention. There are many similarity computation methods, including kernels, distances, and non-metric similarity [70]. The kernel-based similarity adopts the kernel trick to compute similarity and make predictions in a high-dimensional implicit feature space without the explicit coordinates of the transformed data. It can extract various nonlinear relationships and is suitable for processing heterogeneous multi-modal features [71], [72]. There are some kernels that measure the similarity relations between two sets [73], [74]. Thus, this work introduces a general linear kernel [74] to capture the similarity between multi-modal features better and achieve the properties required for multi-modal interactive attention.

The affinity matrix $\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}$, representing the similarity between the main-modal features $\boldsymbol{M}^n$ and the auxiliary-modal features $\boldsymbol{A}^n$, is computed by

$$
\begin{aligned}
&\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n} \\
&= \left[(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})_{ij}\right]_{d_{a^n}\times d_m} \\
&= \begin{bmatrix}
\mathcal{K}(\boldsymbol{A}_1^n,\boldsymbol{M}_1^n) & \cdots & \mathcal{K}(\boldsymbol{A}_1^n,\boldsymbol{M}_j^n) & \cdots & \mathcal{K}(\boldsymbol{A}_1^n,\boldsymbol{M}_{d_m}^n) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\mathcal{K}(\boldsymbol{A}_i^n,\boldsymbol{M}_1^n) & \cdots & \mathcal{K}(\boldsymbol{A}_i^n,\boldsymbol{M}_j^n) & \cdots & \mathcal{K}(\boldsymbol{A}_i^n,\boldsymbol{M}_{d_m}^n) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\mathcal{K}(\boldsymbol{A}_{d_{a^n}}^n,\boldsymbol{M}_1^n) & \cdots & \mathcal{K}(\boldsymbol{A}_{d_{a^n}}^n,\boldsymbol{M}_j^n) & \cdots & \mathcal{K}(\boldsymbol{A}_{d_{a^n}}^n,\boldsymbol{M}_{d_m}^n)
\end{bmatrix} \\
&= \begin{bmatrix}
(\boldsymbol{A}_1^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_1^n & \cdots & (\boldsymbol{A}_1^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_j^n & \cdots & (\boldsymbol{A}_1^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_{d_m}^n \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
(\boldsymbol{A}_i^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_1^n & \cdots & (\boldsymbol{A}_i^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_j^n & \cdots & (\boldsymbol{A}_i^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_{d_m}^n \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
(\boldsymbol{A}_{d_{a^n}}^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_1^n & \cdots & (\boldsymbol{A}_{d_{a^n}}^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_j^n & \cdots & (\boldsymbol{A}_{d_{a^n}}^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_{d_m}^n
\end{bmatrix}
\end{aligned}
\tag{1}
$$

where $\mathcal{K}(\boldsymbol{A}_i^n, \boldsymbol{M}_j^n)$ is a general linear kernel, and $\boldsymbol{W}^n$ is required to be a positive semi-definite matrix according to Definition 3.1. Furthermore, $\boldsymbol{W}^n$ is an adaptive weight to adjust the multi-modal similarity along with the training of the network. Hence, the learning of the general linear kernel in each MIA module is driven by different pairwise modal data, and learning similarity from data is more unbiased than traditional similarity measurement methods [75].

Definition 3.1. *If $\boldsymbol{W}_{c\times c}$ is a positive semi-definite matrix, then the function $k : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ given by $k(x, y) = x^T W y$ is a general linear kernel.*

In addition, the positive semi-definite matrix $\boldsymbol{W}^n$ is a real symmetric matrix, which can be decomposed as an orthogonal matrix $\boldsymbol{P}$ according to Theorem 3.1. More explicitly, the general linear kernel can be expressed by

$$
\begin{aligned}
\mathcal{K}(\boldsymbol{A}_i^n, \boldsymbol{M}_j^n) &= (\boldsymbol{A}_i^n)^{\mathrm{T}}\boldsymbol{W}^n\boldsymbol{M}_j^n \\
&= (\boldsymbol{A}_i^n)^{\mathrm{T}}(\boldsymbol{P}^n)^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{P}^n\boldsymbol{M}_j^n \\
&= (\boldsymbol{P}^n\boldsymbol{A}_i^n)^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{P}^n\boldsymbol{M}_j^n).
\end{aligned}
\tag{2}
$$

This indicates that the feature embeddings $\boldsymbol{A}^n$ and $\boldsymbol{M}^n$ are projected into an orthogonal common space and maintain their norm. Orthogonality is a desirable quality in deep networks, partially because multiplication by an orthogonal matrix leaves the norm of the original matrix unchanged, avoiding the signals vanishing or exploding due to repeated matrix multiplication [76]. Thus, the orthogonality of the general linear kernel makes it feasible to stack multiple MIA modules to add more modalities. With the help of the general linear kernel, the affinity matrix $\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}$ adaptively represents the similarity between the $\boldsymbol{a}^n$ modality and the main modality during training. Moreover, the affinity matrix $\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}$ combines the information of the $\boldsymbol{a}^n$ modality and the main modality.

Theorem 3.1. *Every real symmetric matrix $\boldsymbol{Q}$ can be decomposed as $\boldsymbol{Q} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{P}$, where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with eigenvalues of $\boldsymbol{Q}$ as its diagonal elements.*

Transform. Then, the affinity matrix $\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}$ is encoded into the interactive attention weights $\boldsymbol{P}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})$ for improving the main-modal features. It can be computed by

$$
\boldsymbol{P}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) = \left[\boldsymbol{P}_1(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}), \cdots, \boldsymbol{P}_j(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}), \cdots, \boldsymbol{P}_{d_m}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})\right]
$$

$$
\boldsymbol{P}_j(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) = \begin{bmatrix} \boldsymbol{P}_{j1}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) \\ \vdots \\ \boldsymbol{P}_{ji}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) \\ \vdots \\ \boldsymbol{P}_{jd_{a^n}}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) \end{bmatrix} = \begin{bmatrix} \frac{exp(\mathcal{K}(\boldsymbol{A}_1^n,\boldsymbol{M}_j^n))}{\sum_k^{d_{a^n}} exp(\mathcal{K}(\boldsymbol{A}_k^n,\boldsymbol{M}_j^n))} \\ \vdots \\ \frac{exp(\mathcal{K}(\boldsymbol{A}_i^n,\boldsymbol{M}_j^n))}{\sum_k^{d_{a^n}} exp(\mathcal{K}(\boldsymbol{A}_k^n,\boldsymbol{M}_j^n))} \\ \vdots \\ \frac{exp(\mathcal{K}(\boldsymbol{A}_{d_{a^n}}^n,\boldsymbol{M}_j^n))}{\sum_k^{d_{a^n}} exp(\mathcal{K}(\boldsymbol{A}_k^n,\boldsymbol{M}_j^n))} \end{bmatrix}
\tag{3}
$$

where $\boldsymbol{P}_{ji}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})$ represents the importance of the i-th auxiliary-modal feature vector over the j-th main-modal feature vector. So $\boldsymbol{P}_j(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})$ denotes the importance of all auxiliary-modal feature vectors over the the j-th main-modal feature vector. Next, the interactive attention weights $\boldsymbol{P}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})$ are applied to improve main-modal features $\boldsymbol{M}^n$ by:

$$
\hat{\boldsymbol{M}}_{m\leftarrow a^n}^n = \left[\hat{\boldsymbol{M}}_1^n, \quad \cdots, \quad \hat{\boldsymbol{M}}_j^n, \quad \cdots, \quad \hat{\boldsymbol{M}}_{d_m}^n\right] \tag{4}
$$

$$
\hat{\boldsymbol{M}}_j^n = \textstyle\sum_i^{d_{a^n}} \boldsymbol{A}_i^n \cdot \boldsymbol{P}_{ji}(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n}) \tag{5}
$$

where '$\cdot$' denotes the Hadamard product. Eq. (1), (3), (4) and (5) reveals that $\hat{\boldsymbol{M}}_{m\leftarrow a^n}^n$ are the updated main-modal features augmented by auxiliary-modal features.

Inspired by the works [77], [78], this paper introduces a self-gate mechanism further extract more discriminative features from the updated main-modal features $\hat{\boldsymbol{M}}_{m\leftarrow a^n}^n$ and suppress possible noisy features. The self-gate mechanism can be expressed by:

$$
\boldsymbol{G} = \sigma(\boldsymbol{K}_1 \odot \hat{\boldsymbol{M}}_{m\leftarrow a^n}^n + \boldsymbol{b}_1) \tag{6}
$$

$$\boldsymbol{M}_{gated}^{n} = \hat{\boldsymbol{M}}_{m\leftarrow a^n}^{n} \cdot \boldsymbol{G} \tag{7}$$

where $\odot$ denotes the convolution operation, and σ is logistic sigmoid activation function. $\boldsymbol{K}_1$ and $\boldsymbol{b}_1$ are the filter kernel and the bias, respectively. $\boldsymbol{G}$ is the gating coefficients, and $\boldsymbol{M}_{gated}^{n}$ is the improved main-modal features after a self-gated mechanism. From Eq.(5), each updated main-modal feature vector $\hat{\boldsymbol{M}}_j^n$ is the weighted sum of all auxiliary-modal feature vectors $[\boldsymbol{A}_1^n, \cdots, \boldsymbol{A}_i^n, \cdots, \boldsymbol{A}_{d_{a^n}}^n]$ and the interactive attention weights $\boldsymbol{P}_j(\boldsymbol{S}_{\boldsymbol{M}^n\boldsymbol{A}^n})$. So each updated main-modal feature vector $\hat{\boldsymbol{M}}_j^n$ may be affected by the auxiliary-modal noisy features. The self-gate mechanism utilizes convolution operations to learn more discriminative emotional patterns, and uses the sigmoid function σ to encode these patterns as gating coefficients $\boldsymbol{G}$. The gating coefficients $\boldsymbol{G}$ are multiplied by the updated main-modal features $\hat{\boldsymbol{M}}_{m\leftarrow a^n}^{n}$ themselves to preserve more discriminative emotional features and filter possible noisy features.

Fusion. Since $\boldsymbol{M}_{gated}^{n}$ is the improved features of the main-modal features $\boldsymbol{M}^n$ after a multi-modal interactive attention and a self-gated mechanism, both $\boldsymbol{M}_{gated}^{n}$ and $\boldsymbol{M}^n$ are the main-modal features. It is reasonable to fuse them into the new main-modal features $\boldsymbol{M}^{n+1}$ via some simple homogeneous fusion methods [79], [80], such as concatenation, sum, max, convolution, etc. The convolutional block attention module (CBAM) [81] fused the average-pooled features and max-pooled features by a standard convolution layer as a spatial attention map. Inspired by CBAM, this paper employs convolution to fuse $\boldsymbol{M}_{gated}^{n}$ and $\boldsymbol{M}^n$. The simple fusion process is as follows:

$$\boldsymbol{M}^{n+1} = f_N(\sigma(\boldsymbol{K}_2 \odot [\boldsymbol{M}_{gated}^{n} : \boldsymbol{M}^n] + \boldsymbol{b}_2)) \tag{8}$$

where $\boldsymbol{K}_2$ and $\boldsymbol{b}_2$ are the convolutional kernel and bias, respectively. $f_N(\cdot)$ represents the batch normalization function [82], and $[:]$ means the concatenation operation.

Kernel Regularization Loss. In Eq. (1), the $\mathcal{K}(\boldsymbol{A}_i^n, \boldsymbol{M}_j^n)$ requires the matrix $\boldsymbol{W}^n$ to be a positive semi-definite matrix. This paper designs a kernel regularization loss to satisfy these constraints. Here denotes $\boldsymbol{S} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^n, \cdots, \boldsymbol{W}^N\}$. For positive semi-definite matrix $\boldsymbol{W} \in \boldsymbol{S}$, there is the following definition and theorem:

Definition 3.2. *A matrix $\boldsymbol{W}$ is the positive semi-definite matrix iff,*

1. *$\boldsymbol{W}$ is symmetric.*
2. *$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x} \geq 0,\ \forall \boldsymbol{x} \in \mathbb{R}^c$.*

Theorem 3.2. *For a $c \times c$ matrix $\boldsymbol{W}$, it is a positive semi-definite matrix iff it is symmetric and all its eigenvalues are non-negative.*

Proof. Let a symmetric matrix $\boldsymbol{W}$ have spectral decomposition $\boldsymbol{W} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{P}$ where $\boldsymbol{P}$ is orthogonal. Consider the quadratic form of $\boldsymbol{W}$.

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x} = \overbrace{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}}^{\boldsymbol{y}^{\mathrm{T}}}\boldsymbol{\Lambda}\overbrace{\boldsymbol{P}\boldsymbol{x}}^{\boldsymbol{y}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y} = \sum_i \lambda_i y_i^2 \tag{9}$$

For $\boldsymbol{x} \neq 0$, we have $\boldsymbol{y} \neq 0$ and thus $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x} = \sum_i \lambda_i y_i^2 \geq 0$. Hence $\boldsymbol{W}$ is positive semi-definite matrix. □

According to Definition 3.2 and Theorem 3.2, $\boldsymbol{W}$ should be symmetric $\boldsymbol{W}^{\mathrm{T}} = \boldsymbol{W}$, and all eigenvalues of $\boldsymbol{W}$ should be non-negative $\lambda_i = |\lambda_i|, i = 1, \cdots, c$. In order to satisfy such conditions, the kernel regularization loss can be calculated by:

$$\mathcal{L}_{KRL} = \textstyle\sum_{\boldsymbol{W}\in\boldsymbol{S}}(|\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}| + \sum_i^c |\lambda_i = |\lambda_i||) \tag{10}$$

where $\lambda_1, \cdots, \lambda_c$ are the eigenvalues of $\boldsymbol{W}$. $|\cdot|$ is $\ell 1$ norm, which means the sum of the absolute values of the vector elements. Specifically, the first term in $\mathcal{L}_{KRL}$ constrains $\boldsymbol{W}$ to be a symmetric matrix, and the second term constrains that all the eigenvalues of $\boldsymbol{W}$ are non-negative. The kernel regularization loss $\mathcal{L}_{KRL}$ encourages $\boldsymbol{W}$ to be a positive semi-definite matrix, so that $\mathcal{K}(\boldsymbol{A}_i^n, \boldsymbol{M}_j^n)$ becomes the general linear kernel to learn the interactive relations between the main-modal features $\boldsymbol{M}^n$ and the auxiliary-modal features $\boldsymbol{A}^n$ better.

3.2.2 *Dual residual structure.*

Inspired by the works [55], [83], this paper employs a dual residual connection to avoid vanishing gradients and degradation problems [83], [84] during deep training. Here, $\boldsymbol{M}^1$ is denoted as the initial main-modal features extracted from main-modal data, and $\boldsymbol{M}_{MIAs}$ is denoted as the improved main-modal features of $\boldsymbol{M}^1$ after multiple multi-modal interactive attention modules. This work employs a residual connection around $\boldsymbol{M}_{MIAs}$ and $\boldsymbol{M}^1$, which can be expressed by

$$\begin{aligned}\boldsymbol{M}_{res} &= \boldsymbol{M}^1 + \boldsymbol{M}_{MIAs} \\ &= \boldsymbol{M}^1 + \boldsymbol{f}_{MIAs}(\boldsymbol{M}^1, \boldsymbol{A}^1, \cdots, \boldsymbol{A}^N)\end{aligned} \tag{11}$$

where $\boldsymbol{f}_{MIAs}(\cdot)$ denotes the functions of multiple multi-modal interactive attention modules. $\boldsymbol{M}_{res}$ is the residual main-modal features. Then the second residual connection around a feed-forward network (FFN) is adopted to increase nonlinear representations.

$$\check{\boldsymbol{M}} = \boldsymbol{M}_{res} + f_{FFN}(\boldsymbol{M}_{res}) \tag{12}$$

where $\check{\boldsymbol{M}}$ is the new main-modal features as the output of multi-modal interactive attention network.

3.3 Regression or Classifition

The multi-modal affective analysis tasks include multi-modal sentiment analysis and multi-modal emotion recognition tasks. This work conducts both multi-modal sentiment analysis tasks and multi-modal emotion recognition tasks to demonstrate the generalization of the proposed model.

In the multi-modal sentiment analysis task, the sentiment regression model consists of one fully connected layer and a joint loss. The joint loss is calculated as the weighted sum of the mean absolute error loss and the kernel regularization loss. It is given by

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{MAE} + \gamma\mathcal{L}_{KRL} \tag{13}$$

where γ is a hyper parameter to balance the two losses. $\mathcal{L}_{MAE}$ is a mean absolute error (MAE) loss, which can be expressed as

$$\mathcal{L}_{MAE} = -\frac{1}{N}\sum_{n=1}^{N}|y^{(n)} - z^{(n)}| \tag{14}$$

$$z = \boldsymbol{\theta}^{\mathrm{T}}\check{\boldsymbol{M}} + \boldsymbol{b}_3 \tag{15}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{b}_3$ are the weight and bias of the fully-connected layer, respectively. $\check{\boldsymbol{M}}$ is the output feature of MIA-Net. $y^{(n)}$ and $z^{(n)}$ are the true label and regression value of the n-th sample, respectively.

In the multi-modal emotion recognition task, the emotion recognizer contains a fully-connected layer, a softmax function and a joint loss, which can be expressed as

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{CE} + \gamma\mathcal{L}_{KRL} \tag{16}$$

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N} y_n log(P(\hat{y}_n)) \tag{17}$$

$$P(\hat{y}) = softmax(\boldsymbol{\theta}^{\mathrm{T}}\check{\boldsymbol{M}} + \boldsymbol{b}_3). \tag{18}$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy (CE) loss. N is the number of samples. y_n and $\hat{y}_n$ are the true label and predicted label of the n-th sample. It is worth mentioning that the output of the emotion recognizer is the probability distribution over all emotional categories. The emotion category with the highest probability is selected as the predicted label $\hat{y}$. $P(\hat{y})$ is the probability that the sample is classified as Class $\hat{y}$.

4 Experiments and Results

4.1 Dataset

The proposed method is evaluated on the following datasets.

CMU-MOSI [34] is a multi-modal sentiment analysis dataset. CMU-MOSI dataset has 93 videos from the YouTube website, and these videos are segmented into 2199 opinion utterances. Each utterance is manually labeled with a score between -3 to 3, where -3 denotes the strongest negative and +3 denotes the strongest positive. According to the official settings, we use 1284 utterances for training, 229 utterances for validation, and 686 utterances for testing.

CMU-MOSEI [35] is the largest multi-modal sentiment analysis dataset with more than 22,000 video segments. These video segments contain various themed videos and monologue videos from YouTube. Each segment can be split into acoustic, visual, and textual modalities. Each video segment is annotated with an integer score between -3 (strongly negative) to +3 (strongly positive). Also, this paper performs binary and 7-class sentiment classification on this dataset. Following the official settings, we use 16328 video segments for training, 1872 video segments for validation, and 4663 video segments for testing.

MELD [36] is a multi-modal emotion recognition dataset encompassing audio and visual as well as textual information. MELD dataset contains 13708 utterances from the TV series *Friends*. Each utterance is annotated with an emotional label, including anger, disgust, fear, joy, neutral, sadness, or surprise. For a fair comparison with works [53], [58], MELD dataset is split into training, validation and testing sets with 9955, 1106 and 2607 utterances, respectively.

4.2 Baselines

MIA-Net is compared to various baselines, including typical and state-of-the-art models. As mentioned in Section 2.2, these baselines can be categorized as follows:

Tensor-based fusion models contain TFN [42], LMF [23].

Modality-invariant representation learning models contain MISA [49], MIB [48], MMIM [47].

Attention-based fusion models contain soft attention based model (CM-BERT [51], DialogueRNN [52], A-DMN [53], Multi-modal Routing model [54]), multihead attention model (BBFN [56], SSL [57]), Transformer based model (CTNet [58], Cross-modal Transformer (CT) [20]), sparse attention model (TCM-LSTM [25]), contextual inter-modal attention model (CIM [18]).

Other models, such as a novel quantum-inspired multi-modal fusion framework (QMF) [85], an unidirectional modality translation model MTSA [86] and the heterogeneous graph convolution fusion models (HIS-MSA [87], MM-DFN [88], DSAGCN [89]).

4.3 Experimental Settings

Hyperparameters and implementation. This paper develops MIA-Net on the Fairseq [90] framework. Many modalities can express emotion in affective analysis tasks. This paper takes the modalities from the CMU-MOSI, CMU-MOSEI, and MELD datasets as examples, including text, audio, and video modalities. Therefore, the textual feature dimension d_t, the acoustic feature dimension d_a, and the visual feature dimension d_v are 1024, 768, and 256, respectively. The hyperparameter γ in Eq. (13) is 0.3. In training, this paper adopts Adam optimizer [91] to train MIA-Net, and sets the initial learning rate as $1.7e^{-5}$ and the batch size as 18. This paper trains MIA-Net with a maximum of 50 epochs, and stops training if the validation loss does not decrease for six consecutive epochs.

Evaluation metrics. In CMU-MOSI and CMU-MOSEI datasets, the model performance is evaluated with mean absolute error (MAE), binary accuracy (Acc-2), F1, and seven-class accuracy (Acc-7). F1 score on each category and the weighted average F1 (w-F1) are adopted in the MELD dataset.

4.4 Comparison with State-of-The-Art Approaches

The proposed method is compared with state-of-the-art approaches on the CMU-MOSI, CMU-MOSEI, and MELD datasets. 'MIA-Net (F)' means MIA-Net adopts the proposed feature set described in Section 3.1. 'MIA-Net (F_{cmu})' means MIA-Net adopts the initial feature set released by the team of the CMU-MOSI and CMU-MOSEI datasets, including the COVAREP features [92], Facet features [93], word embedding.

Results on CMU-MOSI dataset. As shown in Table 1, the proposed method under the feature set F and F_{cmu} shows improvement over typical approaches on the seven-class accuracy, the binary accuracy, F1 score, and MAE. Compared with the state-of-the-art model SSL [57], MIA-Net achieves 4.42%, 1.56%, and 1.22% improvement in seven-class accuracy, binary accuracy, and F1 score, respectively. The remarkable improvement in seven-class accuracy proves that MIA-Net effectively fuses multi-modal features and makes full use of multi-modal complementary information to reduce ambiguity in sentiment expression.

Results on CMU-MOSEI dataset. To evaluate MIA-Net's ability in handling more challenging multi-modal

TABLE 1
Performance of MIA-Net against State-Of-The-Art (SOTA) Models on CMU-MOSI Dataset.

	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
QMF [85]	33.53	79.74	79.62	0.92
TCM-LSTM [25]	35.40	81.70	81.80	0.90
MulT [20]	40.00	83.00	82.80	0.87
MISA [49]	42.30	83.40	83.60	0.78
CM-BERT [51]	44.90	84.50	84.50	0.73
BBFN [56]	45.00	84.30	84.30	0.78
MTSA [86]	46.40	86.80	86.80	0.70
MMIM [47]	46.65	86.06	85.98	0.70
HIS-MSA [87]	48.40	86.01	85.99	0.67
MIB [48]	48.60	85.30	85.30	0.71
SSL [57]	48.93	88.23	88.57	0.58
MIA-Net (F)	**53.35**	**89.79**	**89.79**	0.59
MIA-Net (F_{cmu})	52.92	89.02	88.97	**0.55**

TABLE 2
Performance of MIA-Net against State-Of-The-Art (SOTA) Models on CMU-MOSEI Dataset.

	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
QMF [85]	47.80	80.69	79.77	0.64
LMF [94]	49.30	80.80	-	0.62
TCM-LSTM [25]	50.60	81.40	81.60	0.61
Multi-modal Routing [54]	51.60	81.70	81.80	-
MulT [20]	51.80	82.50	82.30	0.58
MISA [49]	52.20	85.50	85.30	0.56
MTSA [86]	52.90	85.50	85.30	0.54
MIB [48]	54.10	85.70	85.70	0.60
MMIM [47]	54.24	85.97	85.94	0.53
HIS-MSA [87]	54.90	83.96	83.89	0.51
BBFN [56]	54.80	86.20	86.10	0.53
MIA-Net (F)	**56.44**	87.93	87.98	**0.49**
MIA-Net (F_{cmu})	56.15	**88.59**	**88.56**	0.49

sentiment analysis tasks, this subsection reports the performance of MIA-Net on the large multi-modal sentiment analysis dataset CMU-MOSEI. The results are presented in Table 2. Table 2 shows that MIA-Net under the feature set F and F_{cmu} achieves the best performance among all the approaches on seven-class accuracy, binary accuracy, F1 score, and MAE. Specifically, MIA-Net outperforms the state-of-the-art method BBFN [56] by 1.64% on seven-class accuracy, 1.73% on binary accuracy, 1.88% on F1 score, and 0.04 on MAE, respectively. It demonstrates MIA-Net's robustness in terms of handling more challenging multi-modal sentiment analysis tasks.

Results on MELD dataset. As presented in Table 3, MIA-Net significantly outperforms other methods. Table 3 also points out that the accuracy varies widely between classes, caused by the data imbalance in MELD dataset. However, the accuracy of the weak categories is greatly improved in MIA-Net. For example, the accuracy of the Fear and Disgust categories respectively increases by 17.40% and 17.08% compared to CTNet [58]. MIA-Net can exploit the vital information of the auxiliary modalities to assist the representation learning of the main modality, rather than treating all modalities equally. Therefore, MIA-Net reduces the useless information and noises from the auxiliary modalities and improves all categories, especially weak ones.

To sum up, MIA-Net reaches excellent performance across multiple datasets, proving its effectiveness and robustness in multi-modal affective analysis.

5 Analysis and Discussions

5.1 Ablation Studies

This subsection designs various ablation studies to analyze the importance of different modalities and MIA-Net. Here, the 'Audio-only' model means only acoustic features are fed to a regressor to predict the sentiment. The 'Text-only' model is a sentiment regressor with only textual features, as does the 'Video-only' model. The'TV-MIA' model and 'TA-MIA' model denote the MIA-Net with text-video modalities and that with text-audio modalities, respectively. 'TAV-MIA' model is the proposed method with text-audio-video modalities. The results of the ablation studies are shown in Table 4.

As can be seen from Table 4, the 'Audio-only' model and the 'Video-only' gain fairly low accuracy. In addition, the 'Text-only' model performs best among unimodal models. This demonstrates that text description contains the dominant sentiment information for sentiment analysis, and the visual and acoustic information contributes less to sentiment than textual semantics. However, the performance improves slightly over the 'Text-only' model when the text modality and one of the other two modalities are available. It proves that the video and audio modalities can provide supplemental information for textual sentiment analysis. These findings coincide with this paper's innovation, which treats the modality that contributes the most to emotion as the main modality and the others as auxiliary modalities.

Furthermore, MIA-Net based on trimodal data significantly outperforms the unimodal and bimodal models in Table 4, indicating the effectiveness of MIA-Net in handling the data with three or more modalities. This result also highlights the necessity of analyzing human sentiment in a multi-modal perspective.

5.2 Generalization to New Datasets, Tasks and Modalities

This subsection offers opportunities to study whether MIA-Net can transfer knowledge between different datasets and between different tasks, and whether it can generalize to different features and more modalities. The results are shown in Table 5 and Table 6.

Transfer on different datasets. In transfer experiments between different datasets, MIA-Net is trained on one dataset and is tested on other datasets. The transfer experiments include the transfer between small and large datasets. The CMU-MOSI dataset and the CMU-MOSEI dataset are multi-modal sentiment analysis datasets with 2199 opinion videos and 22000 themed or monologue videos, respectively. The sample size of the CMU-MOSEI dataset is about ten times larger than that of the CMU-MOSI dataset. Thus, the performance of MIA-Net trained on the CMU-MOSEI dataset achieves a 2.67% gain in seven-class accuracy as compared to the state-of-the-art model SSL [57] on the CMU-MOSI dataset. This demonstrates that MIA-Net can learn the latent sentiment information from the multi-modal

TABLE 3
Performance of MIA-Net against State-Of-The-Art (SOTA) Models on MELD Dataset.

	Netural (%)	Sadness (%)	Anger (%)	Joy (%)	Surprise (%)	Fear (%)	Disgust (%)	Avg (%)
DSAGCN [89]	74.40	22.10	46.90	49.60	45.50	4.80	8.70	58.70
MM-DFN [88]	77.76	22.93	47.82	54.78	50.60	0.00	0.00	59.46
DialogueRNN [52]	77.44	34.59	43.65	54.40	52.51	11.68	7.89	60.25
A-DMN [53]	78.94	24.92	40.98	57.41	55.38	8.60	3.47	60.45
CTNet [58]	77.40	32.50	44.60	56.00	52.70	10.00	11.20	60.50
MIA-Net	**79.83**	**41.92**	**52.52**	**61.98**	**58.68**	**27.40**	**28.28**	**65.82**

TABLE 4
Ablation Study on CMU-MOSI Dataset.

	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
Audio-only	15.60	42.53	25.73	1.47
Video-only	17.35	50.61	49.27	1.51
Text-only	47.81	87.96	87.95	0.60
TA-MIA	48.83	88.72	88.73	0.59
TV-MIA	49.42	88.11	88.14	0.60
TAV-MIA	53.35	89.79	89.79	0.59

sentiment analysis datasets, and achieves superior transfer performance in low-resource datasets. However, when MIA-Net is trained on a small CMU-MOSI dataset and is tested on a large CMU-MOSEI dataset, its performance is similar to Multi-modal Routing model [54] and LMF [94] but lower than BBFN [56]. This is because the sample size of CMU-MOSI dataset is about one-tenth of that of CMU-MOSEI dataset, and MIA-Net does not learn enough on the small CMU-MOSI dataset. This reveals the importance of data amount for deep learning models. Future work will focus on semi-supervised multi-modal learning to address this issue.

Transfer on different tasks. As for transfer experiments between different tasks, MIA-Net is trained on one task, and is tested on another task after fine-turning the last fully-connected layer. These experiments include the transfer between the emotion recognition task and sentiment analysis task. As shown in Tables 5, 1 and 2, MIA-Net, under the transfer learning from emotion classification task to sentiment regression tasks, achieves comparable performance to the state-of-the-art model SSL [57] on CMU-MOSI dataset and the state-of-the-art method BBFN [56] on CMU-MOSEI dataset. When performing transfer learning from sentiment regression task to emotion recognition task, MIA-Net increases by 4.18 % compared to CTNet [58] on MELD dataset. These results imply that MIA-Net learns more generalizable multi-modal features in transfer learning across different tasks.

Generalization on different features. This subsection also performs the MIA-Net under different feature sets on CMU-MOSI and CMU-MOSEI datasets, including the feature set F described in Section 3.1 and the initial feature set F_{cmu} released by the team of CMU-MOSI and CMU-MOSEI datasets. The feature set F_{cmu} includes the COVAREP features [92], Facet features [93], and RoBERTa features based on the input word embedding. The feature set F contains the RoBERTa features based on VQ-Wav2Vec embeddings, Fabnet features, and RoBERTa features based on the input word embedding. It can be observed in Table 6 that MIA-Net under the feature set F_{cmu} has a similar performance to MIA-Net under the feature set F on both CMU-MOSEI and CMU-MOSI datasets, which indicates that MIA-Net can generalize to different features.

Generalization on more modalities. This subsection conducts experiments to verify the performance of MIA-Net in trimodal or more modal tasks. It is worth noting that the CMU-MOSEI, CMU-MOSI, and MELD datasets only have three modalities. To simulate datasets with more than three modalities, this experiment converts audio and video modalities into two new modalities through a fully connected layer and adds them to MIA-Net. The results are reported in Table 7. As the number of modalities increases from 2 to 3, the performance of MIA-Net improves significantly on CMU-MOSI dataset. This illustrates that the multi-modal comprehensive information can reduce the ambiguity of emotion expression and achieves superior performance. Comparing the MIA-Net based on three modalities and based on five modalities, they have similar performance. MIA-Net based on five modalities does not get much improvement due to the limited amount of information, since the new additional modalities come from the original audio and video modalities. More importantly, the results on three modalities and five modalities indicate that MIA-Net can maintain efficient training when stacking multiple MIA modules to expand more modalities.

To sum up, MIA-Net's weights learned from one multi-modal affective dataset/task generalize well to a new multi-modal affective dataset/task. Moreover, MIA-Net can support different feature sets and different tasks based on trimodal and multi-modal data. These results demonstrate the generalization of MIA-Net.

5.3 Discussion on the Main and Auxiliary Modalities

Since different modalities contribute differently in emotion expression, it is rational to design a multi-modal framework to treat them unequally. In MIA-Net, the modality that contributes the most to emotion is regarded as the main modality and the other modalities as the auxiliary modalities. MIA-Net adopts multi-modal interactive attention modules to make important information flow from the auxiliary modalities to the main modality, which improves the discriminative performance of main-modal representation. Therefore, this subsection conducts the multi-modalities swapping experiments to verify the effectiveness of MIA-Net in terms of assistive fusion. The assistive fusion means a process where the performance of one modality is improved by the stimulation in another modality, which refers to the weakly coupled cross-modal learning [95].

TABLE 5
Transfer Performance of MIA-Net between Different Datasets and Tasks.

Transfer Settings	Training Dataset	Test Dataset	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)	w-F1 (%)
Large dataset → Small dataset	CMU-MOSEI	CMU-MOSI	51.60	88.57	88.55	0.58	-
Small dataset → Large dataset	CMU-MOSI	CMU-MOSEI	49.31	83.80	83.92	0.58	-
Classification Task → Small Regression Task	MELD	CMU-MOSI	47.09	89.02	89.00	0.63	-
Classification Task → Large Regression Task	MELD	CMU-MOSEI	55.24	86.25	86.39	0.51	-
Regression Task → Classification Task	CMU-MOSEI	MELD	-	-	-	-	64.68

TABLE 6
The Performance of MIA-Net under Different Feature Sets.

Dataset	Feature	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
CMU-MOSEI	F	56.44	87.93	87.98	0.49
	F_{cmu}	56.15	88.59	88.56	0.49
CMU-MOSI	F	53.35	89.79	89.79	0.59
	F_{cmu}	52.92	89.02	88.97	0.55

TABLE 7
The MIA-Net Performance of Generalization on More Modalities in CMU-MOSI Dataset

The number of modality	CMU-MOSI			
	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
N=2	48.83	88.72	88.73	0.59
N=3	53.35	89.79	89.79	0.59
N=5	52.62	89.79	89.73	0.57

TABLE 8
Multi-modalities Swapping Experiments in CMU-MOSEI and MELD Datasets

$\boldsymbol{m}$	$\boldsymbol{a}^1$	$\boldsymbol{a}^2$	CMU-MOSEI				MELD
			Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)	w-F1 (%)
Text	Video	Audio	56.40	87.29	87.18	0.49	65.23
Text	Audio	Video	56.44	87.93	87.98	0.49	65.82
Audio	Video	Text	53.03	85.95	85.89	0.53	59.71
Video	Audio	Text	54.29	85.97	86.02	0.53	58.48

TABLE 9
The Unimodal Performance in CMU-MOSEI and MELD Datasets

Modality	CMU-MOSEI				MELD
	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)	w-F1 (%)
Audio	42.28	70.46	69.05	0.78	31.43
Audio [49], [88]	41.34	61.37	59.80	0.82	42.72
Video	41.03	66.42	62.65	0.80	31.38
Video [49], [88]	41.30	60.70	59.96	0.84	32.34
Text	52.70	86.94	86.91	0.54	61.32
Text [49], [88]	51.70	84.22	84.15	0.56	56.95

Table 8 shows the swapping experiments between different main modalities $\boldsymbol{m}$ and between the first auxiliary modality $\boldsymbol{a}^1$ and the second auxiliary modality $\boldsymbol{a}^2$.

In Table 8, the results of swapping the position of audio and video in auxiliary modality are almost identical. Since MIA-Net selects the important information from the auxiliary-modal features one by one to improve the main-modal feature representation, each auxiliary-modal feature is independent of each other in MIA-Net. Thus, swapping the position of different auxiliary modalities has little impact on affective analysis performance.

Table 8 presents that the results of swapping different main modalities vary differently. The MIA-Net with text as the main modality works best, and others perform poorly. The reason can be found in the comparison of Table 8 and Table 9. In Table 9, the performance of the unimodal model based on audio and video is poor. It is speculated that audio and video modalities contribute lower than text modality in emotion expression [25], [27], and they may contain much noise in the data pre-processing or feature extraction process [25]. Thus, the performance of MIA-Net with audio and video as the main modalities is lower than that of MIA-Net with text as the main modality. It is reasonable to treat the modality that contributes the most to emotion as the main modality and the other modalities as the auxiliary modalities. It is worth noting that, comparing the MIA-Net with audio as the main modality in Table 8 with the audio model in Table 9, it can be found that MIA-Net extracts textual and visual emotional information to assist the audio model as much as possible and greatly improves it. So as does in video modality. These results demonstrate the effectiveness of the multi-modal interactive attention module in terms of assistive fusion.

5.4 Discussion on the Impact of γ

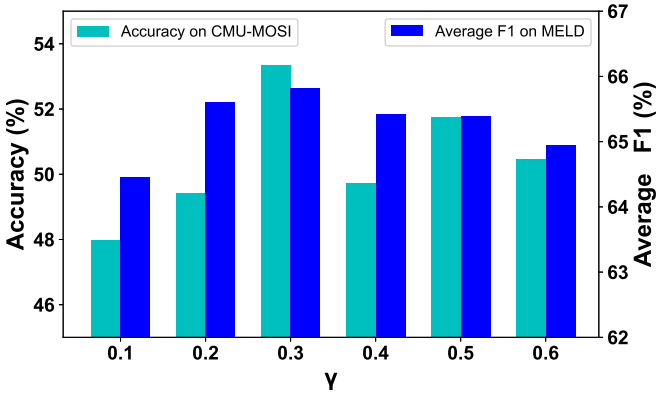

Fig. 2. Comparison of MIA-Net with Different Parameter γ on CMU-MOSI and MELD Datasets

In MIA-Net, the hyperparameter γ is crucial in balancing the MAE loss / cross-entropy loss and kernel regularization loss. This subsection conducts experiments on the CMU-MOSI and MELD datasets to analyze the impact of γ changing incrementally from 0.1 to 0.6.

As shown in Fig. 2, as γ increases starting from 0.1, the performance of MIA-Net improves significantly. This

further verifies that kernel regularization loss gradually increases its contribution to affective analysis and improves the performance as the value of γ increases. In addition, the performance gains plateaus as γ reaches 0.3, and the model does not benefit greatly from increasing the value of γ. It is suspected that MAE loss or cross-entropy loss is influenced by the ratio of kernel regularization loss and gradually reduces the contribution to affective analysis, affecting the regression or classification performance. Therefore, the performance peaks at $\gamma = 0.3$ and then begins to decline gradually.

5.5 Discussion on the Impact of Joint Loss

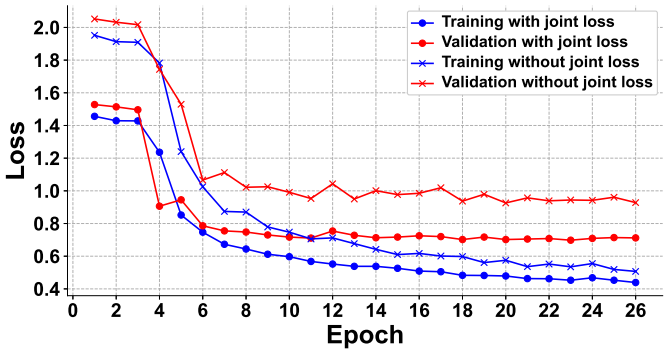

Fig. 3. The Ioint Loss of MIA-Net during Training and Validating on CMU-MOSI Dataset

To further validate the effectiveness of the joint loss, this subsection conducts a control experiment on the joint loss in CMU-MOSI dataset. Fig. 3 shows the results with joint loss and without joint loss during training and validation. It is worth noting that the joint loss is the joint supervision loss of MAE loss and kernel regularization loss. Here, MIA-Net without joint loss means MIA-Net only with MAE loss.

As shown in Fig. 3, the joint loss during training and validation decreases rapidly with increasing epochs and converges in the last few epochs. This demonstrates that MIA-Net has a good learning pattern under the joint supervision of MAE loss and kernel regularization loss. In Fig. 3, the joint loss maintains a considerable advantage over the MAE loss in validation, even though the MAE loss gradually tends towards the joint loss in training. This illustrates that the training can quickly converge to a better local optima under the guidance of the joint loss, further verifying the effectiveness of the joint loss.

5.6 Discussion on Performance and Efficiency of different Fusion Techniques

In order to verify that the proposed fusion strategy is indeed effective, this subsection compares it with other fusion strategies in terms of performance and efficiency on the MELD dataset. The fusion strategies include tensor-based fusion methods (TFN [42]) and attention-based fusion methods (CIM [18], multi-head attention (MHA) [55], Transformer [55], Cross-Transformer (CT) [20]). In these experiments, the feature fusion part of the framework uses different fusion methods, and the rest is kept identical. The experimental results are shown in Fig. 4 and Table 10.

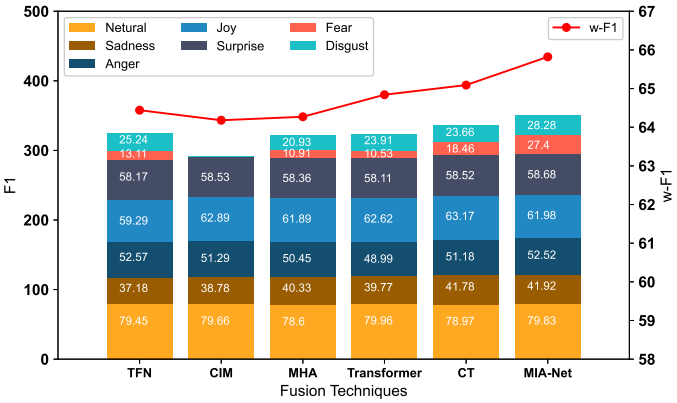

Fig. 4. Comparison of MIA-Net with Different Fusion Techniques on MELD Dataset

TABLE 10
Comparison of Existing Fusion Techniques in Terms of the Floating-Point Operations (FLOPs) and the Number of Parameters (Params) under Trimodal and Four-modal Data

Fusion Techniques	Trimodal data		Four-modal data	
	FLOPs (M)	Params (M)	FLOPs (M)	Params (M)
TFN	137.97	22.99	9011.20	1467.39
CIM	37.80	6.30	51.95	8.66
MHA	75.63	12.60	108.70	22.31
Transformer	176.39	54.61	240.95	81.12
CT	729.91	159.52	1116.16	238.49
MIA-Net	42.10	9.33	48.94	9.33

Performance Comparison. TFN reaches impressive results in Fig. 4, but it requires far more computational resources in training due to its high-dimensional property. Transformer model captures bimodal dependencies and handles the high-dimensional features well, thus outperforming TFN. In addition, it achieves better performance than MHA, because Transformer model is composed of several MHA and takes full advantage of MHA. However, MIA-Net realizes the effective multi-modal fusion via a novel assistive fusion strategy and significantly outperforms the tensor-based and transformer-based methods, as shown in Fig. 4.

Efficiency Comparison. The floating-point operations (FLOPs) and the number of parameters (Params) are popular complexity measures of neural networks. Table 10 shows a comparison of FLOPs and Params for different fusion techniques when supporting trimodal and four-modal data. Here, the units are million (M) for Param, and Mega (M) for FLOPs.

As shown in Table 10, TFN model requires approximately sixty times more FLOPs and parameters to generalize from trimodal data to four-modal data. This is due to the outer product of each high-dimensional modal feature and the following huge fully connected layer.

In attention-based fusion methods, Transformer is composed of multiple MHA, and a cross-modal Transformer (CT) for each pairwise modalities consists of three Transformers. Therefore, CT has more FLOPs and parameters than Transformer, followed by MHA. However, when MHA,

Transformer and CT extend one more modality in the form of auxiliary structures (that is, all modalities are mapped to a modality), it still requires nearly 1.5 times more FLOPs and parameters. If follows the extension method in work [20], it will lead to large growth in parameters and FLOPs as the number of modalities increases, which limits the generalization of the model.

Table 10 reports that MIA-Net has relatively few FLOPs and parameters when supporting trimodal and four-modal data. In addition, MIA-Net only requires a small linear increase in FLOPS and keeps stable parameter counts to generalize to more modal data. This is because the number of FLOPs is mainly concentrated in each lightweight MIA module, and the parameters are concentrated in a feed-forward network (FFN) after the last MIA module. Thus, MIA-Net enables quick generalization to trimodal or multi-modal data under linear computation and stable parameter counts.

Therefore, MIA-Net improves the trade-off between performance and efficiency, and achieves satisfactory performance and impressive efficiency.

6 Conclusion

This paper proposes a novel multi-modal interactive attention network for multi-modal affective analysis. MIA-Net is a general multi-modal fusion model, which enables quick generalization to trimodal or multi-modal data under linear computation and stable parameter counts. Specifically, MIA-Net consists of multiple multi-modal interactive attention modules. Each multi-modal interactive attention module adaptively captures important information from the auxiliary modalities and adds them to assist the main-modal representation. Moreover, MIA-Net can support N-modal affective analysis task through stacking $N-1$ lightweight multi-modal interactive attention modules, and maintains impressive performance. Extensive experiments on multiple benchmark datasets verify the effectiveness and generalization of MIA-Net.

Future research will focus on building a large-scale multi-modal emotion database and semi-supervised multi-modal learning models.

Acknowledgments

The authors wish to acknowledge the authors of CMU-MOSI dataset [34], CMU-MOSEI [35], MELD dataset [36], and the pretained models [64], [65], [68] used in this work.

References

[1] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, vol. 149, pp. 102 447–102 473, 2020.

[2] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2020.

[3] R. Francese and P. Attanasio, "Supporting depression screening with multimodal emotion detection," in *ACM International Conference Proceeding Series*, 2021, pp. 1 – 8.

[4] T. L. Praveena and N. Lakshmi, "Multi label classification for emotion analysis of autism spectrum disorder children using deep neural networks," in *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*, 2021, pp. 1018 – 1022.

[5] S. Sarabadani, L. C. Schudlo, A. A. Samadani, and A. Kushski, "Physiological detection of affective states in children with autism spectrum disorder," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 588–600, 2020.

[6] M. Dahmane, J. Alam, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, "A multimodal non-intrusive stress monitoring from the pleasure-arousal emotional dimensions," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[7] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 4, pp. 1052 – 1068, 2004.

[8] Y. Choi, J. Wiebe, and R. Mihalcea, "Coarse-grained +/-effect word sense disambiguation for implicit sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 471–479, 2017.

[9] T. Ma, H. Rong, Y. Hao, J. Cao, Y. Tian, and M. Al-Rodhaan, "A novel sentiment polarity detection framework for chinese," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 60–74, 2022.

[10] T. Zhang, X. Gong, and C. L. P. Chen, "Bmt-net: Broad multitask transformer network for sentiment analysis," *IEEE Transactions on Cybernetics*, pp. 1–12, 2021.

[11] C. Singh, S. Wibowo, and S. Grandhi, "A deep learning approach for human face sentiment classification," in *Proceedings - 2021 21st ACIS International Semi-Virtual Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD-Winter 2021*, 2021, pp. 28 – 32.

[12] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 544–550, 2021.

[13] T.-R. Huang, S.-M. Hsu, and L.-C. Fu, "Data augmentation via face morphing for recognizing intensities of facial emotions," *IEEE Transactions on Affective Computing*, pp. 1–9, 2021.

[14] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[15] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 177–188, 2021.

[16] Z. Huang and J. Epps, "An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 653–668, 2020.

[17] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, pp. 107 676–107 688, 2022.

[18] M. S. Akhtar, D. S. Chauhan, and A. Ekbal, "A deep multi-task contextual attention framework for multi-modal affect analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 3, pp. 1–27, 2020.

[19] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 371–378.

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, pp. 6558—-6569.

[21] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[22] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114.

[23] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2247–2256.

[24] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019, pp. 12 136–12 145.

[25] S. Mai and S. Xing, "Analyzing multimodal sentiment via acoustic- and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1424 – 1437, 2021.

[26] F. Huang, K. Wei, J. Weng, and Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 16, no. 3, pp. 1–19, 2020.

[27] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 481 – 492.

[28] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Poczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *33rd AAAI Conference on Artificial Intelligence*, 2019, pp. 6892 – 6899.

[29] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1424 – 1437, 2021.

[30] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end asr models," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7149–7153.

[31] R. Xia, J. Jiang, and H. He, "Distantly supervised lifelong learning for large-scale social media sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 480–491, 2017.

[32] C. Singh, S. Wibowo, and S. Grandhi, "A deep learning approach for human face sentiment classification," in *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, 2021, pp. 28–32.

[33] Y. Li, K. Zhang, J. Wang, and X. Gao, "A cognitive brain model for multimodal sentiment analysis based on attention neural networks," *Neurocomputing*, vol. 430, pp. 159 – 173, 2021.

[34] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[35] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "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 (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[36] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536.

[37] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[38] M. Paleari and C. L. Lisetti, "Toward multimodal fusion of affective cues," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, 2006, pp. 99 – 108.

[39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[40] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[41] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2018.

[42] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[43] S. Wang, L. Hao, and Q. Ji, "Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1084–1097, 2020.

[44] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, United states, 2019, pp. 1–20.

[45] S. Verma, C. Wang, L. Zhu, and W. Liu, "Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis," in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2019-August, 2019, pp. 3627 – 3634.

[46] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1144–1149.

[47] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 9180 – 9192, 2021.

[48] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Transactions on Multimedia*, pp. 1 – 13, 2022.

[49] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[51] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 521––528.

[52] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[53] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, pp. 1–15, 2020.

[54] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 1823 – 1833.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017, pp. 5999 – 6009.

[56] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 6 – 15.

[57] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, Shanghai, China, 2020, pp. 3755 – 3759.

[58] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[59] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based label set generation for multi-modal multi-label emotion detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. Association for Computing Machinery, 2020, pp. 512—-520.

[60] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5301–5311.

[61] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information fusion*, vol. 53, pp. 209–221, 2020.

[62] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9–35, 2019.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[64] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *8th International Conference on Learning Representations, ICLR 2020*, pp. 1–15, 2020.

[65] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020, pp. 1–12.

[66] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[67] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[68] O. Wiles, A. Sophia Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," in *British Machine Vision Conference 2018, BMVC 2018*, 2019, pp. 1–15.

[69] H. An, D. Wu, and Z. Li, "Hybrid self-interactive attentive siamese network for medical textual semantic similarity," *ACM International Conference Proceeding Series*, pp. 52 – 56, 2020.

[70] K. Rieck and P. Laskov, "Linear-time computation of similarity measures for sequential data," *Journal of Machine Learning Research*, vol. 9, pp. 23 – 46, 2008.

[71] M. Engin, L. Wang, L. Zhou, and X. Liu, "Deepkspd: Learning kernel-matrix-based spd representation for fine-grained image recognition," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11206 LNCS, pp. 629 – 645, 2018.

[72] S. Yu, L.-C. Tranchevent, X. Liu, W. Glanzel, J. A. Suykens, B. De Moor, and Y. Moreau, "Optimized data fusion for kernel k-means clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1031 – 1039, 2012.

[73] B.-C. Xu, K. M. Ting, and Z.-H. Zhou, "Isolation set-kernel and its application to multi-instance learning," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 941 – 949.

[74] L. A. Belanche Muñoz, "Developments in kernel design," in *ESANN 2013 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 24-26 April 2013*, 2013, pp. 369–378.

[75] Z. Kang, C. Peng, and Q. Cheng, "Kernel-driven similarity learning," *Neurocomputing*, vol. 267, pp. 210 – 219, 2017.

[76] A. Brock, T. Lim, J. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1031 – 1039, 2017.

[77] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2019.

[78] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632.

[79] M. A. Rahman and R. Laganière, "Mid-level fusion for end-to-end temporal activity detection in untrimmed video." in *BMVC*, 2020, pp. 1 – 13.

[80] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2091 – 2106, 2022.

[81] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, pp. 3 – 19, 2018.

[82] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[84] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, pp. 1–16.

[85] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Information Fusion*, vol. 65, pp. 58–71, 2021.

[86] B. Yang, B. Shao, L. Wu, and X. Lin, "Multimodal sentiment analysis with unidirectional modality translation," *Neurocomputing (Amsterdam)*, vol. 467, pp. 130–137, 2022.

[87] Y. Zeng, Z. Li, Z. Tang, Z. Chen, and H. Ma, "Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis," *SSRN*, pp. 1 – 37, 2022.

[88] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations," vol. 2022-May, 2022, pp. 7037 – 7041.

[89] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis," *Neurocomputing*, vol. 501, pp. 629 – 639, 2022.

[90] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019, pp. 1–6.

[91] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

[92] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," 2014, pp. 960 – 964.

[93] E. L. Rosenberg and P. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.

[94] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, "Low rank fusion based transformers for multimodal sequences," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 29 – 34.

[95] N. M. Seel, *Encyclopedia of the Sciences of Learning*. Springer Science & Business Media, 2011.

Shuzhen Li received the B.S. degree in electronic information engineering from Shantou University, at Guangdong, China, in 2017, and the M.S. degree in electronic and information engineering from South China University of Technology, at Guangzhou, China, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering from South China University of Technology, at Guangzhou, China.

Her research interests include multi-modal affective computing, speech emotion recognition and machine learning.

Tong Zhang (S'12-M'16) received the B.S. degree in software engineering from Sun Yat-sen University, at Guangzhou, China, in 2009, and the M.S. degree in applied mathematics from University of Macau, at Macau, China, in 2011, and the Ph.D. degree in software engineering from the University of Macau, at Macau, China in 2016. Dr. Zhang currently is a professor with the School of Computer Science and Engineering, South China University of Technology, China.

His research interests include affective computing, evolutionary computation, neural network, and other machine learning techniques and their applications. Dr. Zhang is an Associate Editor of the IEEE Transactions on Computational Social Systems. He has been working in publication matters for many IEEE conferences.

Bianna Chen received the B.S. degree in information engineering from Guangdong University of Technology, Guangzhou, China, in 2017, the M.S degree in electronic and information engineering from South China University of Technology, Guangzhou, China, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering from South China University of Technology, Guangzhou, China.

Her research interests include Graph neural network and affective computing.

C. L. Philip Chen (S'88-M'88-SM'94-F'07) received the M.S. degree from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 1985 and the Ph.D. degree from the Purdue University in 1988, all in electrical and computer science.

He is the Chair Professor and Dean of the College of Computer Science and Engineering, South China University of Technology. He is the former Dean of the Faculty of Science and Technology. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE) and European Academy of Sciences and Arts (EASA). He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is also a highly cited researcher by Clarivate Analytics in 2018, 2019, 2020, and 2021.

He was the Editor-in-Chief of the IEEE Transactions on Cybernetics (2020-2021) after he completed his term as the Editor-in-Chief of the IEEE Transactions on Systems, Man, and Cybernetics: Systems (2014-2019), followed by serving as the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013. Currently, he serves as an deputy director of CAAI Transactions on AI, an Associate Editor of the IEEE Transactions on AI, IEEE Trans on SMC: Systems, and IEEE Transactions on Fuzzy Systems, an Associate Editor of China Sciences: Information Sciences. He received Macau FDCT Natural Science Award three times and a First-rank Guangdong Province Scientific and Technology Advancement Award in 2019. His current research interests include cybernetics, computational intelligence, and systems.