Applied Intelligence (2024) 54:3040–3057
https://doi.org/10.1007/s10489-024-05329-w

A multimodal fusion-based deep learning framework combined with local-global contextual TCNs for continuous emotion recognition from videos

Congbao Shi[1] · **Yuanyuan Zhang**[1] · **Baolin Liu**[1]

Accepted: 7 February 2024 / Published online: 21 February 2024

Abstract
Continuous emotion recognition plays a crucial role in developing friendly and natural human-computer interaction applications. However, there exist two significant challenges unresolved in this field: how to effectively fuse complementary information from multiple modalities and capture long-range contextual dependencies during emotional evolution. In this paper, a novel multimodal continuous emotion recognition framework was proposed to address the above challenges. For the multimodal fusion challenge, the Multimodal Attention Fusion (MAF) method is proposed to fully utilize complementarity and redundancy between multiple modalities. To tackle temporal context dependencies, the Local Contextual Temporal Convolutional Network (LC-TCN) and the Global Contextual Temporal Convolutional Network (GC-TCN) were presented. These networks have the ability to progressively integrate multi-scale temporal contextual information from input streams of different modalities. Comprehensive experiments are conducted on the RECOLA and SEWA datasets to assess the effectiveness of our proposed framework. The experimental results demonstrate superior recognition performance compared to state-of-the-art approaches, achieving 0.834 and 0.671 on RECOLA, 0.573 and 0.533 on SEWA in terms of arousal and valence, respectively. These findings indicate a novel direction for continuous emotion recognition by exploring temporal multi-scale information.

Keywords Continuous emotion recognition · Local/global contextual temporal convolutional network · Multi-modal attention fusion · Temporal multi-scale information

1 Introduction

Automatic emotion recognition assumes a vital role in the development of natural and friendly human-computer interaction applications, which enables computers to have higher and more comprehensive intelligence. In recent years, driven by the robust advancement of artificial intelligence and deep learning techniques [1, 2], emotion recognition has been widely applied in various real-life domains, such as vehicle driving [3], healthcare [4], education [5]. Early research in the field of emotion recognition primarily concentrated on the recognition of discrete emotional states through various modalities, including facial expressions, audio cues, and texts [6]. The discrete emotion model proposed by Ekman is used in most studies to express affective states, which contains seven basic emotions such as happy and angry[7]. Nonetheless, as in-depth research in this field has progressed, it has become increasingly apparent that the discrete emotion model exhibits three distinct deficiencies in effectively expressing affective states [8, 9]: (1) The affective states expressed by the discrete emotion model are inherently limited and there exist cultural differences; (2) It is difficult to measure and deal with the correlation between emotion categories; (3) The emotional evolution process of emotion generation, development and transformation cannot be described. As a result, there has been a shift away from discrete emotion recognition towards the prediction of affective states in the continuous dimensional space [10]. A widely

✉ Baolin Liu
liubaolin@ustb.edu.cn

Congbao Shi
G20208826@xs.ustb.edu.cn

Yuanyuan Zhang
zhangyuanyuan@ustb.edu.cn

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, People's Republic of China

embraced dimensional emotion model within this context is the valence-arousal model, proposed by Russell in 1980 [11]. It utilizes the two fundamental dimensions: valence, representing the positive and negative aspects of human emotion, and arousal, indicating the degree of excitement and depression within the emotional scope. Therefore, continuous emotion recognition task, grounded in the dimensional emotion model, aims to devise methodologies capable of effectively predicting subtle and complex affective states.

Two key issues of how to effectively fuse complementary information from different modalities and capture the long-range context dependences remain unresolved in continuous emotion recognition. They not only curtail the robustness and accuracy of the emotion recognition system, but also hinder its broader application in various life scenarios. Theoretically, fusing information from multiple modalities can substantially enhance the recognition accuracy. In most emotion recognition studies [12], feature-level and decision-level fusion strategies have been widely adopted. Feature-level fusion methods generally obtains the fused features by concatenating feature vectors from multiple modalities. The authors of [13] extracted high-level representations from video and audio signals separately, which were then concatenated as input to a Long Short-Term Memory (LSTM) model. Although this strategy is easy to follow and comprehend, it may still be susceptible to the curse of dimensionality. Decision-level fusion approaches feed the features extracted from each modality into separate networks to generate initial predictions, which are then fed into a subsequent recognition model to forecast the final affective states [14, 15]. Although decision-level fusion approaches avoids the dimensionality issue caused by feature concatenation, they may overlook the intricate and nonlinear relationships between different modalities, thereby amplifying the complexity of model training. Furthermore, both of the aforementioned fusion strategies fall short in their ability to explore and exploit modality-specific information and common information among modalities. Recently, the transformer [16] has ignited extensive discussions regarding the application of attention mechanism across diverse research fields [17–19]. In the realm of multi-modal fusion, attention mechanism is proved to be an effective solution for capturing dynamic interactions between different modalities. In response to this challenge, a multi-modal fusion approach utilizing the attention mechanism and the modal interaction matrix is proposed to explore intra-modal and inter-modal information interactions.

In recent years, abundant investigations have shown that modeling temporal context dependencies is beneficial to improving the prediction performance of emotion dimensions. This is because the target dimension values are continuous and the time interval between two adjacent predictions is short. Conventional regression models such as Support Vector Regression (SVR) and Relevance Vector Machine (RVM) were commonly adopted to predict continuous affective states in early continuous emotion recognition research [8, 12]. However, these approaches predict the affective state independently at each time step, lacking the capability to model temporal dependencies. Many methods also relied on Recurrent Neural Networks (RNNs) to capture temporal contextual information and achieving emotion recognition [20, 21]. However, RNNs and LSTMs generally perform poorly in learning very long-range contextual information. Although this kind of networks can theoretically handle sequence data with arbitrary length, it generally suffers from the limitations of gradient vanishing and explosion. Three-Dimensional Convolutional Neural Network (3D CNN) is an extension of CNN in the temporal dimension, which is specifically designed to capture temporal contextual information in consecutive frames. Although some progress has been made in continuous emotion recognition [22, 23], the methods based on 3D CNN incur high computational consumption when capturing long-range contextual information. Therefore, how to effectively model the long-range contextual dependencies in sequential data remains a challenge. To address this problem, the local and global contextual temporal convolutional networks are proposed to explicitly integrate multi-scale contextual information in continuous emotion recognition. Our proposed method is designed on the basis of the Temporal Convolutional Network (TCN) which has been applied to several temporal modeling tasks with exciting results [24, 25]. It allows the receptive field on the time axis to grow exponentially without introducing too many parameters. Furthermore, to the best of our knowledge, few works have explicitly explored the impact of multi-scale information in time dimension on capturing long-term context dependencies.

In summary, continuous emotion recognition is a very challenging task. On the one hand, it requires that the redundancy and complementarity of modal information needs to be fully considered when designing the fusion method. An inappropriate fusion method may reduce robustness of the recognition system. On the other hand, inability to effectively model temporal dynamic information during emotion evolution hinders the improvement of recognition performance. Therefore, a novel multi-modal continuous emotion recognition framework was proposed that simultaneously addresses the challenges of multi-modal data fusion and temporal contextual modeling. For this paper, the main contributions are as follows:

(1) Regarding the issue of multimodal information fusion, a model level fusion method based on attention mechanism has been proposed, which includes intra modal attention and inter modal attention, effectively learn-

ing complex nonlinear relationships between different modalities while promoting dynamic interaction of emotional information.

(2) Regarding the issue of temporal context dependency, the local contextual temporal convolutional network and the global contextual temporal convolutional network have been proposed which combine temporal convolution networks with multi-scale context in time axis to progressively integrate multi-scale temporal contextual information from input feature sequences.

(3) An end-to-end trainable model framework has been built that fuses multi-modal data to predict affective states at the frame level. Experimental results conducted on the RECOLA and SEWA datasets demonstrate the superiority of our proposed framework over the state-of-the-art methods for continuous emotion recognition.

The remainder of this paper is arranged as follows. In Section 2 a brief overview of the related studies has been given. Section 3 describes the details of the proposed architecture for multi-modal continuous emotion recognition. The RECOLA dataset and SEWA dataset, the experimental setup, as well as the experimental results, are reported in Section 4. Conclusion and discussion are presented in Section 5.

2 Related work

2.1 Multi-modal continuous emotion recognition

A growing body of works have recently been devoted to exploring continuous emotion recognition, in which the multi-dimensional space constructed by the dimensional emotion model can express a wide range of subtle and complex affective states. Many continuous emotion recognition studies are based on three benchmark datasets (SEMAINE [26], RECOLA [27], SEWA [28]) used by the Audio/Visual Emotion Challenge (AVEC). The models proposed in these works all focus on how to effectively fuse multiple modalities and how to capture temporal context dependencies, which are crucial for continuous emotion recognition.

For multi-modal fusion, the fusion strategies adopted in most recent works are feature-level fusion and decision-level fusion. In multi-modal continuous emotion recognition using feature-level fusion strategy, features from different modalities are usually concatenated into the multi-modal feature vector, which is then fed into the prediction network to get the final results. Huang et al. [29] combined different modality features in the front-end for training SVR or RVM, enabling simultaneous prediction of valence and arousal values. However, this fusion approach is prone to suffer from dimensional disasters. Dung et al. [30] concatenated the visual and sound features extracted by a dual-stream auto-encoder and fed them into a LSTM to obtain the final prediction results. Although the model was able to learn discriminative emotional features from different modalities, the complementary information in these features was not effectively fused. There are also many multi-modal continuous emotion recognition models based on decision-level fusion. Deng et al. [31] first combined unimodal features into a multi-modal feature vector to train a Deep Bidirectional Long Short-Term Memory network (DBLSTM), and then fed the outputs of multiple unimodal and multi-modal DBLSTMs into a subsequent DBLSTM model to obtain the final prediction results. However, the fusion method ignores the complex nonlinear relationships between different modalities and increases the complexity of model training. Pei et al. [32] explored and proposed a LSTM-based model-level fusion method for audio-visual continuous emotion recognition. This method took into account the complementarity and redundancy among multiple streams from different modalities and utilized an Adaptive Weight Network (AWN) to adequately integrate auxiliary factors such as gender. However, it ignored the influence of other factors in modal information such as head pose and facial occlusion on emotion recognition performance. Schoneveld et al. [33] proposed a model-level fusion method using LSTM to fuse the visual and sound features extracted respectively by the distilled visual feature extraction network and the fine-tuned VGGish backbone network. The advantage of this work was to introduce knowledge distillation technique to emotion recognition to make full use of unlabeled data to improve recognition performance. However, it did not effectively exploit the non-linear interactions between them when fusing these modalities and considering temporal dynamics.

For temporal modeling, Recurrent Neural Networks are adopted as the main structure to capture temporal contextual information in most continuous emotion recognition works. Mao et al. [34] proposed a three-stage model based on the Bidirectional Long Short-Term Memory network (BLSTM) to hierarchically learn emotional context information from video sequences. The model made full use of long-term contextual information in the input data by learning temporal information at different stages. However, the method proposed in this work did not model the temporal dynamics between modalities which was benefit to improve recognition performance. In [35], the authors combined a two-stream network with the Gated Recurrent Unit (GRU) model to learn spatio-temporal representations from video sequences in an end-to-end manner and achieved promising results. It was difficult for this method to explicitly characterize these expressions. 3D CNNs have also been used to extract temporal contextual information from sequential data. For example, the authors of [22, 36] employed 3D convolution networks to consider both temporal motion and spatial appearance

of facial image sequences, thus constructing the network architecture using spatio-temporal information of consecutive frames to improve recognition accuracy. However, these methods based on 3D CNN incurred high computational consumption when capturing long-range contextual information. Therefore, it was difficult for these methods to exploit ultra-long-range context dependencies in videos.

In this paper, a novel framework for continuous emotion recognition has been presented. A model-level fusion method is proposed to fuse multiple feature streams extracted from audio-visual modalities. Additionally, within this framework, the temporal convolutional networks are improved to capture ultra-long-range temporal dependencies.

2.2 Attention mechanisms

To fully exploit the complementary information from different data sources, attention-based fusion models have been widely explored in a large number of applications, such as video question answering [17], video classification [37] and emotion recognition [18, 38, 39]. Wu et al. [40] proposed a novel model framework to fuse facial expressions, head pose and eye gaze information for continuous emotion recognition. Firstly, the extracted head pose, eye gaze cues and facial expression features were fused in the head pose-eye gaze enhanced attention module. Secondly, the guided attention module was designed to adaptively adjust the effect of its noise on facial expression information. However, it was difficult to explore the nonlinear interaction relationship between these features through concatenation operation. In [41], the authors explored a variety of attention fusion strategies and found that the proposed cross-modal hierarchical self-attentional fusion approach was more advantageous after comparisons. The fusion method took features from three modalities as input, namely video, audio, and text, and used self-attention to combine each modality with the remaining modalities and then fused them with features from other modalities in an attentional manner. Although they were able to achieve good performance by understanding inter-modal connections, intra-modal correlations were not well explored in these fusion approaches of above works. In [42], to fully explore the complementarity of visual and aural modalities in video data, authors proposed a joint cross-attention model. It exploited complementary relationships to extract salient features across audio-visual modalities, allowing more accurate predictions of valence and arousal. However, the proposed fusion method did not explore the impact of more auxiliary information in the videos, such as head pose and eye movement.

In this paper, an attention-based model-level fusion method is proposed to efficiently learn complex interactions among multiple modalities. The intra-modal attention modules are used to recalibrate input feature streams so as to highlight salient emotional features. In the inter-modal attention module, the complex interactions between different modalities are learned by cross-modal attention, and the modal interaction matrix is introduced to enhance its ability. Therefore, this fusion approach can simultaneously focus on both the specific emotion information within the modality and the common emotion information between modalities for continuous emotion recognition.

2.3 Dilated convolution and temporal convolutional networks

Although RNNs such as LTSM or GRU are commonly used for sequence modeling tasks to capture temporal information, Temporal Convolutional Networks (TCNs) have received increasing attention in recent research works [24, 43]. In [43], the authors utilized dilated convolutions, causal convolutions and stacked residual blocks to build the general TCN architecture capable of making predictions with information from the far past. After the systematic comparison in a wide range of sequence modeling tasks, they concluded that convolution networks should be considered as a starting point for research in sequence modeling tasks [43]. In [24], the authors proposed a multi-stage architecture for temporal action segmentation. They used dilated convolutions instead of temporal pooling to increase the temporal receptive field, where each stage consisted of a set of dilated convolutions to generate initial predictions that were further refined as inputs to the next stage. Inspired by the stacked hourglass network, Du et al. [44] designed the Temporal Hourglass Convolutional Neural Network (TH-CNN) to capture long-term dynamic dependencies of emotional continuous changes. It established contextual relationships by integrating low-level encoding and high-level decoding information, while a novel supervised strategy, namely Temporal Intermediate Supervision (TIS), was proposed to guide TH-CNN for learning semantic representations in a coarse-to-fine-grained manner. However, the approach did not highlight key information and remove useless and redundant information when integrating low-level coding and high-level decoding information. Hu et al. [25] proposed a two-stage spatio-temporal attention temporal convolution network for video-based continuous dimension emotion recognition. The spatio-temporal attention branch introduced in each stage of the model helped the network to learn different attention levels and adaptively focus on informative spatio-temporal features. At the same time, a smooth loss function was introduced in the training phase to penalize outlier predictions in consecutive frames. However, the model did not well explore temporal context information at different scale. He et al. [45] proposed an adversarial discriminative temporal convolutional network

and an encoder-TCN to perform emotion recognition based on EEG signals. Although the proposed method maintained the representation invariance of features by learning joint temporal information, it failed to effectively reduce the intra-class distance and expand the inter-class distance.

In this paper, a sequence modeling scheme based on temporal convolutional network [43] was proposed to efficiently capture temporal contextual information in sequential data. In contrast to the aforementioned literature, multi-scale contextual temporal information is explicitly integrated by the proposed approach. It includes Local Contextual Temporal Convolutional Network (LC-TCN) and Global Contextual Temporal Convolutional Network (GC-TCN). LC-TCN is composed of dilated convolutions and channel attention to obtain local multi-scale context dependencies in each feature stream, while the GC-TCN module composed of dilated convolutions and dense connections is responsible for obtaining global multi-scale contextual dependencies to predict dimensional affective states. Therefore, it can progressively acquire temporal contextual information, which is crucial for continuous emotion recognition. It should be noted that this sequence modeling scheme can be applied not only to multi-modal continuous emotion recognition, but also to other sequence modeling tasks that take multiple feature sequences as input.

3 method

3.1 Approach overview

In this section, we elaborate on the proposed framework architecture, which utilizes multiple input feature streams from different modalities to predict affective states in the frame-level manner, as shown in Fig. 1. We explore the optimal structure of each part of the framework, which is divided into four parts. The first part is responsible for extracting emotional features from audios and videos respectively. The second part is to feed each modal feature stream into its respective LC-TCN network to capture local contextual information. The third part is to input multiple modal feature streams into the multi-modal attention fusion module to obtain the fused multi-modal features. The fourth part is to input the fused feature stream to the GC-TCN network to get the final prediction results. In the feature extraction part, the handcrafted manner and deep-learning technique are employed to extract informative emotional features from audio and video modalities, respectively. For video modality, geometric features and high-level visual features are extracted as visual emotion representations. The 18-layer Residual Network (ResNet-18) pre-trained on the ImageNet dataset is directly adopted to construct the feature extractor

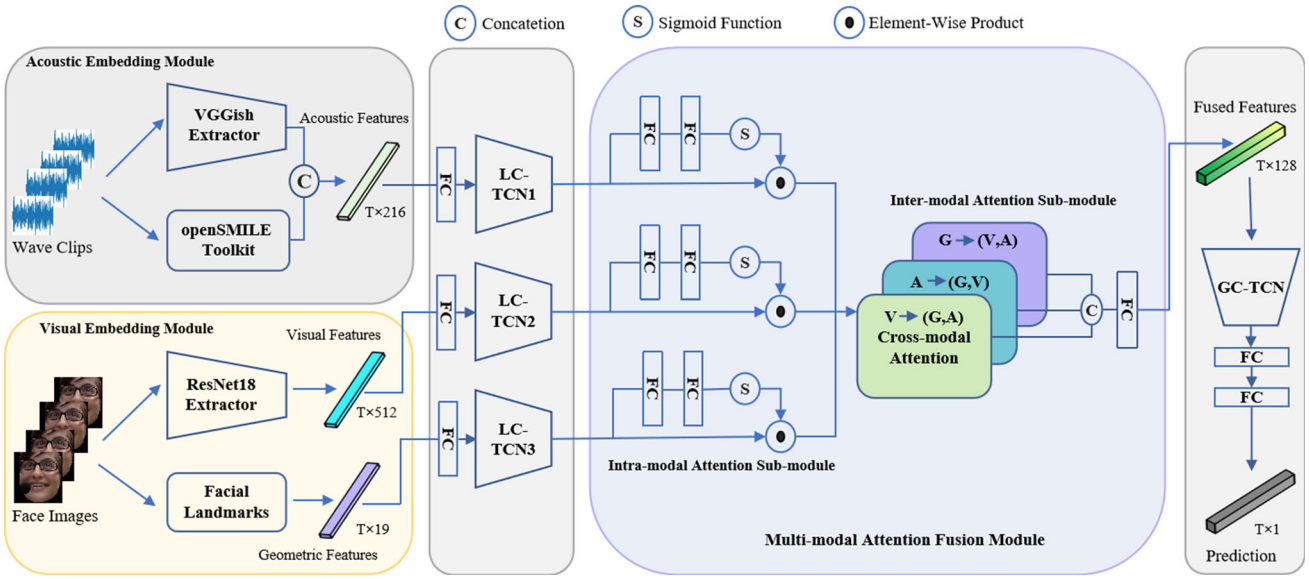

Fig. 1 The overall architecture of the proposed model for continuous emotion recognition. The input feature streams including high-level visual features, geometric features and sound features are extracted from the audio-visual embedding modules. Then, each feature stream is fed into the corresponding LC-TCN to capture local multi-scale context information in the time axis. The multi-modal attention fusion module (MAF) takes these feature streams as input, and by intra-modal attention and inter-modal attention, the complementary emotional information from audio-visual modalities is effectively fused. The GC-TCN is responsible for capturing the global multi-scale temporal contextual information from the fused multi-modal features and feeds them into the prediction sub-network composed of two fully connected layers to predict arousal/valence values

by removing the last fully connected layer of the model to extract high-level visual features from the aligned face image sequence. For audio modality, handcrafted features defined by Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [46] and VGGish-learned sound features are extracted as auditory emotion representations. The VGGish model pre-trained on the AudioSet dataset is adopted to extract features from the log-mel spectrogram as a complement to the eGeMAPS feature set. The concatenation of handcrafted and CNN-learned sound feature vectors is then regarded as the unified representation of acoustic modality. In the local context aggregation part, each feature stream is first fed to the linear hidden layer to standardize into the same dimension. By applying parallel dilated convolutional layers and channel attention to each feature stream, local multi-scale contextual information is integrated and captured through multiple LC-TCNs. In this part, the ability of each feature stream to convey emotional information is improved by the aggregation of local contextual information through the LC-TCN network. In multi-modal fusion part, the proposed multi-modal attention fusion apporach is used to fuse multiple feature streams from different modalities so as to effectively aggregate complementary information for continuous emotion recognition, as shown in Fig. 1. In the dimensional emotion prediction part, the fused multi-modal features are fed into the GC-TCN to capture long-range contextual information during emotion evolution. Meanwhile, considering the subtlety and continuity of emotional evolution, this network is devised in such a way that the deeper layers of the network do not lose too much detailed information. Finally, the arousal values or valence values are predicted by a subsequent prediction sub-network consisting of two fully connected layers. By the way, the overall framework is designed to be end-to-end trainable.

3.2 Local contextual TCN

We first briefly describe the Temporal Convolutional Network (TCN) proposed in [43]. Due to its simple architecture and strong memory ability, it has been gradually used in recent years for sequence prediction and modeling tasks to capture long-range temporal information. Temporal convolution is essentially the one-dimensional convolution operation performed in time axis. And the dilation rate is usually combined with the convolution kernel to expand the receptive field, which is also called dilated convolution. Temporal convolutional network is composed of multiple temporal convolution layers. Therefore, it can take an input sequence with arbitrary length and map it to an output sequence of the same length like the RNNs. In particular, each time step is updated synchronously by a fixed-length time period rather than sequentially. In order to observe the past and future information at each time step like BLSTM, non-causal temporal convolution is employed in this paper.

As shown in Fig. 2, the proposed LC-TCN is mainly composed of multiple local temporal convolution blocks, and the local temporal convolution block is composed of dilated convolutions ($Dilated\_Conv$) with different dila-

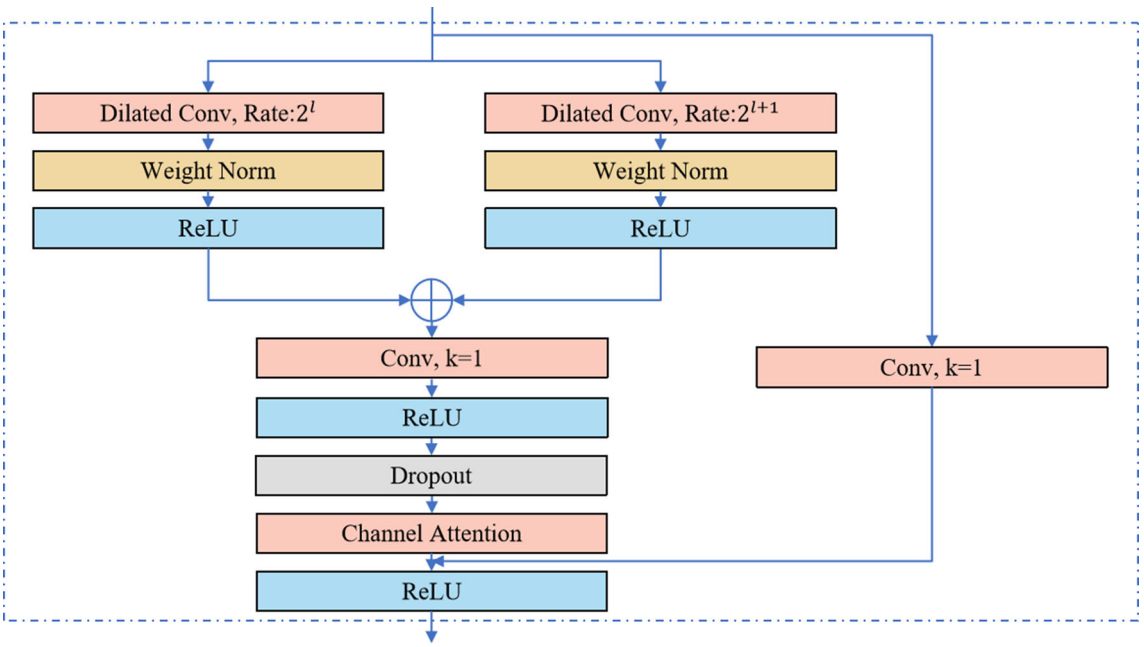

Fig. 2 The architecture of local temporal convolution block. The number of the temporal convolution block located in LC-TCN is denoted as l. For convenience, Squeeze-and-Excitation (SE) attention is selected as channel attention. The kernel size and dilation rate of dilated convolution are 3 and 2^l(or 2^{l+1})

tion factors, a 1D convolution ($Conv1d$) with kernel size of 1, and channel attention ($Channel\_Attn$). In order to describe in detail the execution process of the local timing convolution block, we denote the $l\text{-}th$ local temporal convolution block as TCB_l. Its input and output are respectively $X_l \in \mathbb{R}^{T\times C_0}$ and $Y_l \in \mathbb{R}^{T\times C_1}$, T represents the temporal dimension, and C_0 and C_1 represent the input channel and output channel, respectively. Firstly, the feature sequence X_l is processed by two parallel dilated convolutional layers to compute the local contextual features $DC_{k_1,r_1} \in \mathbb{R}^{T\times C_0}$ and $DC_{k_2,r_2} \in \mathbb{R}^{T\times C_0}$. The output of main branch $Y_l^{'}$ is then calculated by 1D convolutional layer and channel attention module with the concatenation of two local contextual features as input. Finally, $Y_l^{'}$ is point-wise added by the output of residual branch to obtain the final feature sequence Y_l. The residual connection is introduced to reduce the training difficulty. The specific calculation process mentioned above is presented as follows:

$$\begin{aligned} Y_l^{'} &= TCB_l\left(X_l\right) \\ &= Channel\_Attn\left(Conv1d\left(DC_{k_1,r_1} \oplus DC_{k_2,r_2}\right)\right), \\ DC_{k,r} &= Weight\_Norm\left(Dilated\_Conv_{k,r}\left(X_l\right)\right) \end{aligned} \tag{1}$$

$$Y_l = Conv1d\left(X_l\right) + Y_l^{'} \tag{2}$$

where $\oplus$ represents concatenation operation, k and r denote the kernel size and expansion rate of the dilated convolution, respectively. In the local temporal convolution block, a 1D convolution layer with kernel size 1 is adopted to integrate the local multi-scale context information extracted by the parallel dilated convolutional layers, and keep the same output dimension. Note that in addition to the modules mentioned above, the ReLU layer is placed after each convolutional layer to calculate activation values, and the Dropout layer is used to alleviate model overfitting.

Subsequently, the Squeeze-and-Excitation (SE) module proposed to obtain the best performance is selected as the channel attention layer to adaptively recalibrate the channel features. In the module, average pooling operation ($AvgPool$) is utilized to integrate emotional information in the temporal dimension of feature sequence X. The descriptor is fed into two standard convolutional layers ($Conv2d$) with kernel size of 1, and the channel attention map generated by sigmoid function is element-wise multiplied by original input X. The execution process of the module is described as:

$$Channel\_Attn\left(X\right) = \sigma\left(Conv2d\left(Conv2d\left(AvgPool\left(X\right)\right)\right)\right) \odot X \tag{3}$$

where σ denotes the sigmoid function and the symbol $\odot$ indicates element-wise multiplication. Our LC-TCN is designed to obtain a large range of receptive field by increasing the expansion factor per layer as the network goes deeper. This network is used to explicitly learn rich affective patterns in multi-scale contexts for each input feature stream.

3.3 Multi-modal attention fusion module

As illustrated in Fig. 1, the proposed fusion module consists of two parts: the intra-modal attention sub-module and the inter-modal attention sub-module. After the input feature stream is processed by the network described in subsection 3.2, the module first recalibrates the different features within the modality according to their importance to highlight effective emotional features. Then, the inter-modal attention module composed of multiple parallel cross-modal attentions is used to efficiently fuse complementary information from different modalities. Furthermore, the modal interaction matrix is introduced to facilitate this sub-module so as to learn complex interactions between different modalities and their correlations.

3.3.1 Intra-modal attention sub-module

To highlight salient emotional features in each modality, specific intra-modal attention modules are performed separately on each input feature stream containing local temporal contextual information. Referring to [47], two fully connected layers (FC) and a sigmoid activation function are utilized to calculate the attention weights of intra-modal features and element-wise multiply them with the original input feature vector, which is formulated as:

$$Y_t = X_t \odot \sigma\left(\left(X_t W_1\right) W_2\right),\ t = 1,\ 2,\ 3,\ \cdots,\ T \tag{4}$$

where $\odot$ indicates element-wise multiplication, X_t and Y_t denote the feature vector of the $t\text{-}th$ time step in an input feature sequence and the corresponding output vector, respectively. W_1 and W_2 represent the parameters of the first fully connected layer and the second fully connected layer, respectively. σ represents the sigmoid function. Note that we reduce the output channels of the first fully connected layer.

3.3.2 Inter-modal attention sub-module

Inspired by the multi-head attention module proposed in transformer [16], we propose the inter-modal attention module to capture complementary emotional information in multiple modalities. As shown in Fig. 1, the module is composed of multiple parallel cross-modal attention branches, which essentially use the scaled dot product attentions on queries (Q), keys (K) and values (V). To concretely describe the general structure of this module, let us denote the feature vector at time step t of the $i\text{-}th$ feature sequence from multiple modalities as Z_t^i. Therefore, the feature vectors of multiple input sequences at time step t can be denoted as $Z_t = \left(Z_t^1, Z_t^2, Z_t^3, \cdots, Z_t^m\right)$, $m \in [1, M]$, $t \in [1, T]$, where m indicates the number of input modalities. Here,

it is assumed that $m = 3$ to facilitate subsequent elaboration. In one of the attention branches, the query vector is computed from the feature vector Z_t^1, while key and value vectors are computed from the other feature vectors. Firstly, the three feature vectors Z_t^1, Z_t^2, Z_t^3 are projected to the same dimension d_{model} through linear layers, respectively. Then, the query $Q_t^1 \in \mathbb{R}^{d_q}$ is computed by the Z_t^1 and the learnable weights $W_q \in \mathbb{R}^{d_{model} \times d_q}$, d_q represents the dimension of the query vector, that is

$$Q_t^1 = Z_t^1 W_q \tag{5}$$

Inspired by [48] and [49], We refer to the matrix obtained by matrix multiplication of Z_t^2 and Z_t^3 as the Modal Interaction Matrix (MIM), which can enhance the module's ability to learn complex interactions between different modalities. The remaining two eigenvectors Z_t^2, Z_t^3 are first projected into several different subspaces, and then the key K_t^1 and value V_t^1 are calculated according to the above process which are defined as follows:

$$K_t^1 = \left(\left(Z_t^2 W_k^2\right)^T * Z_t^3 W_k^3\right) \oplus \left(\left(Z_t^2 W_k^2\right)^T * Z_t^3 W_k^3\right)^T \tag{6}$$

$$V_t^1 = \left(\left(Z_t^2 W_v^2\right)^T * Z_t^3 W_v^3\right) \oplus \left(\left(Z_t^2 W_v^2\right)^T * Z_t^3 W_v^3\right)^T \tag{7}$$

where $*$ denotes matrix multiplication and $\oplus$ denotes concatenation operation, $W_k^2 \in \mathbb{R}^{d_{model} \times d_k}$, $W_k^3 \in \mathbb{R}^{d_{model} \times d_k}$, $W_v^2 \in \mathbb{R}^{d_{model} \times d_v}$, $W_v^3 \in \mathbb{R}^{d_{model} \times d_v}$ are learnable weight parameters, d_k and d_v denote feature dimensions of key and value, respectively.

The cross-modal attention branch that obtains the query vector based on Z_t^1 is shown in (8) and Fig. 3. At each time step t of input feature sequences, we compute the query Q_t^1 and key K_t^1, then compute the attention weights via dot-product attention and apply them to the value V_t^1 so as to get the final output Z_t^o. In the experiment, d_{model}, d_q, d_k and d_v are set to the same dimension. The rest of the parallel cross-modal attention branches calculated in the same way as this branch. The difference is that the corresponding query vector is calculated using other feature vector, and the remaining branches have their own learnable weight parameters. Finally, the output values of each parallel branch are concatenated and fed into the linear layer to obtain the final values. In this module, complementary emotional information is aggregated in the form of multiple parallel attention branches with different modal features as query vectors.

$$\begin{aligned} Z_t^o &= Attention_{z^1 \to (z^2, z^3)}\left(Z_t^1, Z_t^2, Z_t^3\right) \\ &= softmax\left(\frac{Q_t^1 {K_t^1}^T}{\sqrt{d_k}}\right) V_t^1 \end{aligned} \tag{8}$$

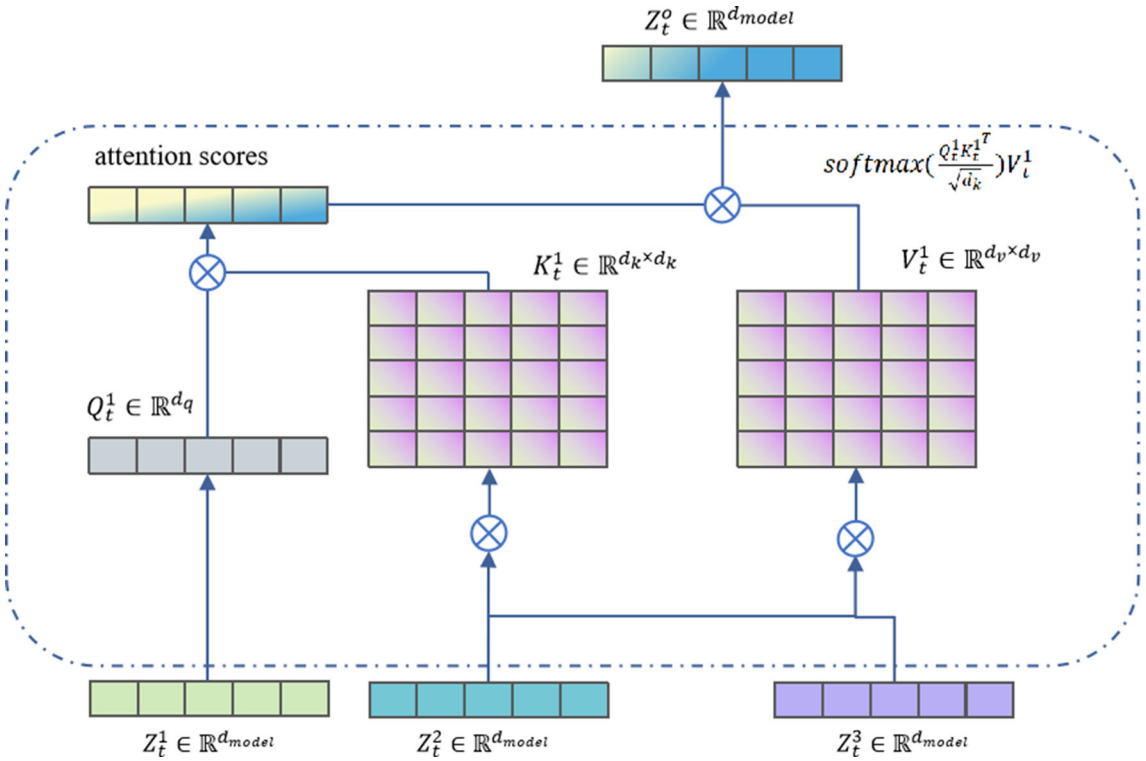

Fig. 3 The architecture of cross-modal attention module $z^1 \to (z^2, z^3)$. The symbol $\otimes$ denotes matrix multiplication. The uni-modal feature vectors (z^1, z^2, z^3) are fed into the module after the standardization linear layers

3.4 Global contextual TCN

In this subsection, the Global Contextual Temporal Convolutional Network (GC-TCN) designed by us is also built on the temporal convolutional network, as shown in Fig. 4. However, the network differs fundamentally from the LC-TCN proposed in subsection 3.2, its goal is to obtain global multi-scale contextual temporal information from the fused multi-modal feature sequence in order to continuously predict affective states. To achieve this goal, we consider the design of the network structure in three ways, which are also different from the LC-TCN. Firstly, GC-TCN has a deeper network structure and a larger convolution kernel to expand the receptive field in the time axis, and can utilize temporal contextual information with the sequence length of ultra-long time steps. Secondly, when designing the global temporal convolutional block, we consider that deeper temporal convolutional layers will gradually lose the ability to perceive subtle changes during emotion evolution. Although it contains two parallel dilated convolution branches, where the expansion rate of one branch increases gradually with the depth of the network, while the expansion rate of the other branch gradually decreases with the depth of the network. Finally, the dense connection structure [50] is introduced into the network to integrate multi-scale temporal features from different layers in a denser manner and improve its robustness.

The GC-TCN is formed by stacking multiple global temporal convolutional blocks. The block is composed of two parallel the dilated convolutions ($Dilated\_Conv$), the 1D convolution ($Conv1d$) and the weight normalization layer ($Weight\_Norm$). The $l\text{-}th$ global temporal convolutional block is denoted as GCB_l. As shown in (9), the output feature sequence O_l is calculated by the concatenation of two global contextual features which are generated by two parallel dilated convolution layers with the feature sequence F_l as input.

$$\begin{aligned} O_l &= GCB_l\left(F_l\right) \\ &= Conv1d\left(DC_{k_1,r_1} \oplus DC_{k_2,r_2}\right), \\ DC_{k,r} &= Weight\_Norm\left(Dilated\_Conv_{k,r}\left(F_l\right)\right) \end{aligned} \tag{9}$$

where $F_l \in \mathbb{R}^{T\times C}$ and $O_l \in \mathbb{R}^{T\times C}$ are respectively the input and output of GCB_l, the symbol $\oplus$ denotes concatenation operation, k and r indicate the kernel size and dilation rate of the dilated convolution. Due to the dense connection structure, the input of GCB_l is computed by the sum of all outputs of the global temporal convolutional blocks before this block and original input feature sequence F_0. The calculation process is defined as follows:

$$F_l = F_0 + \sum_{i=1}^{l} O_i \tag{10}$$

3.5 Affective dimensional prediction sub-network

To predict the arousal or valence dimension, the high-level emotion representation learned by the proposed GC-TCN is fed into the prediction sub-network. It consists of two fully connected layers and a non-linear activation layer in the middle. For the convenience of understanding and description,

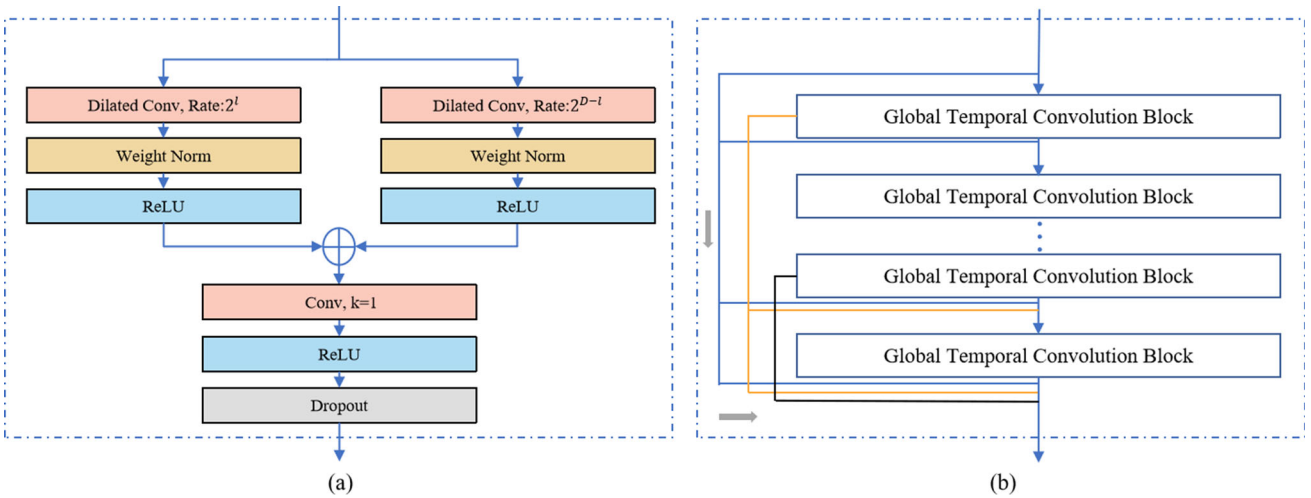

Fig. 4 The architecture of GC-TCN. (a) displays the global temporal convolution block, which contains two parallel dilated convolution branches and a 1D convolutional layer. The GC-TCN learns long-range multi-scale context information in time axis by stacking these blocks. (b) illustrates that the GC-TCN which is composed of dense connections and global temporal convolution blocks. The symbols l and D denote the layer number of the block and total blocks in GC-TCN. The kernel size and dilation rate of dilated convolution are 7 and 2^l(or 2^{D-l})

algorithm 1 summarizes the overall framework forward pass process after feature extraction.

Algorithm 1 the Proposed framework forward pass

Input: $\mathbf{Z}^1, \mathbf{Z}^2, \cdots, \mathbf{Z}^{\mathbf{m}} \in V$, the number of feature sequences from modalities m, the set of feature sequences V. We use $m = 3$ in our experiments.

Output: $\mathbf{P} \in [-1, +1]^{length(\mathbf{Z}) \times 1}$, the prediction values of model on arousal or valence dimension.

Parameters: the framework includes all of the following parameters:
$\mathbf{W_I^1}$, $\mathbf{W_I^2}$, $\mathbf{W_I^3}$, $\mathbf{b_I^1}$, $\mathbf{b_I^2}$, $\mathbf{b_I^3}$, the parameters of standardization linear layers before LC-TCNs.
$\mathbf{W_{LCTCN}^1}$, $\mathbf{W_{LCTCN}^2}$, $\mathbf{W_{LCTCN}^3}$, parameters of each LC-TCN; $\mathbf{W_{GCTCN}}$, the GC-TCN parameters.
$\mathbf{W_{AMA}^1}$, $\mathbf{W_{AMA}^2}$, $\mathbf{W_{AMA}^3}$, parameters of each intra-modal attention sub-module; $\mathbf{W_{RMA}}$, inter-attention sub-module parameters.
$\mathbf{W_{f1}}$, $\mathbf{W_{f2}}$, $\mathbf{b_{f1}}$, $\mathbf{b_{f2}}$, the parameters of the final two fully connected layers.

/* LC-TCN modules to extract local context information from three feature sequences */

$l \longleftarrow length(\mathbf{Z})$
for $i = 1, 2, 3$ do
$\mathbf{Z_L^i} \longleftarrow RELU\left(\mathbf{W_I^i Z^i} + \mathbf{b_I^i}\right)$
$\mathbf{Z_L^i} \longleftarrow LCTCN_i\left(\mathbf{Z_L^i} | \mathbf{W_{LCTCN}^i}\right)$
end for

/* Multi-modal attention module to fuse complementary information from different modalities */

for $i = 1, 2, 3$ do
for $t \in \{1, 2, 3, \ldots, l\}$: $\mathbf{e_t^i} \longleftarrow IntraModalAttention_i\left(\mathbf{Z_L^i}[t] | \mathbf{W_{AMA}^i}\right)$
$\mathbf{Z_I^i} \longleftarrow \left\{\mathbf{e_1^i}, \mathbf{e_2^i}, \mathbf{e_3^i}, \ldots, \mathbf{e_l^i}\right\}$
end for
for $t \in \{1, 2, 3, \ldots, l\}$ do
$\mathbf{f_t} \longleftarrow InterModalAttention\left(\mathbf{Z_I^1}[t], \mathbf{Z_I^2}[t], \mathbf{Z_I^3}[t] | \mathbf{W_{RMA}}\right)$
end for
$\mathbf{Z_F} \longleftarrow \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_l\}$

/* GC-TCN module to extract global context information from the fused feature sequence */

$\mathbf{Z_G} \longleftarrow GCTCN\left(\mathbf{Z_F} | \mathbf{W_{GCTCN}}\right)$
return $\mathbf{P} = \mathbf{W_{f2}} RELU\left(\mathbf{W_{f1}} \mathbf{Z_G} + \mathbf{b_{f1}}\right) + \mathbf{b_{f2}}$

4 Experiments and results

To verify the accuracy and effectiveness of our proposed multi-modal continuous emotion recognition model, extensive experiments are conducted on the RECOLA dataset and SEWA dataset adopted by the Audio/Visual Emotion Challenge (AVEC) and fair comparisons are made with the state-of-the-art methods. The datasets, implementation details, evaluation metric and experimental results will be described at length in subsequent subsections.

4.1 Datasets and features extraction

REmote COLlaborative and Affective interaction (RECOLA) [27] is a widely adopted and open multi-modal continuous emotion dataset in the field, which was constructed to study the socio-emotional behavior of multi-modal data under remote collaboration tasks. The dataset simultaneously recorded multiple modal data, including videos, audios and physiological signals (electrocardiogram, electrodermal activity), generated by the natural interactions of 27 French-speaking subjects. These recordings were continuously annotated by six French-speaking evaluators in two emotional dimensions of arousal and valence, and then all labels were resampled at a frame rate of 40 ms. The dataset was equally divided into three parts: training, validation and testing. Each part contains 9 recorded samples, and each recording is 5 minutes long with a total of 7501 frames. The visualization of sample label is shown in Fig. 5. It should be noted that factors such as gender and age of subjects are balanced in this division to avoid unnecessary effects on emotion recognition.

SEWA dataset [28] consists of audiovisual recordings of participants' spontaneous actions captured using the field recording paradigm, and contains data for both audio and video modalities. Discussions of subjects from German (DE), Hungarian (HU) and Chinese (CN) after watching a set of commercials were recorded via a dedicated video chat platform. The duration of the recordings varied from 40 seconds to 3 minutes. These recordings were continuously annotated at every 100 milliseconds by evaluators on three emotion dimensions of arousal, valence, and liking. To facilitate comparison with other methods, this paper is conducted using the dataset of AVEC2019 challenge [51], which is a subset of the SEWA database. The AVEC2019 dataset was divided into three parts: training set, validation set and test set. These three parts contain 68, 28 and 104 records respectively. Chinese subjects only exist in the test set and the labels of test set are not publicly available. The training and validation sets contain 34 and 14 German and Hungarian subject samples, respectively. For the convenience of comparison, the samples of subjects from the part of Hungarian culture are chosen to carry out the experiment.

In our experiments, we utilize visual and acoustic baseline features extracted from the AVEC2016 challenge [52] which adopt the RECOLA as emotion analysis corpus. Regarding visual features, the 316-dimensional geometric features provided by AVEC2016 are employed, and the calculation process is described as follows. 49 facial landmarks were firstly traced from each frame and then aligned with the average shape from standard points. The 196-dimensional features were calculated from the difference between the landmark positions in the previous and current video frame and the difference between the coordinates of the aligned landmarks and the coordinates of the mean shape. The 49-dimensional features were obtained by computing the Euclidean distance between the median of the standard landmarks and each aligned landmark in the video frame. The

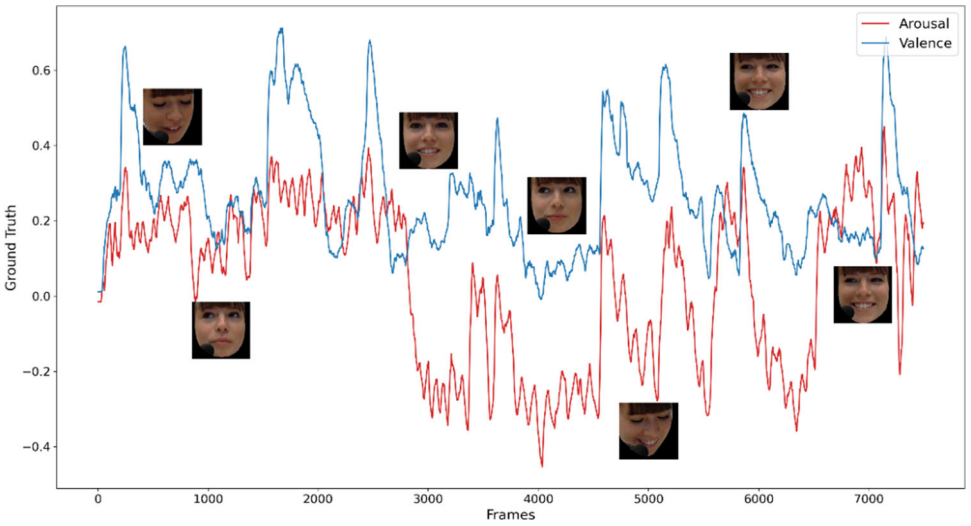

Fig. 5 The visualization of sample label. The horizontal axis represents the frame rate. The emotional dimension value of the vertical axis is [-1,1]

remaining 71-dimensional features correspond to Euclidean distances and angles between points in three different groups. On this basis, we reduce the feature dimensionality by obtaining 19-dimensional features representing 99% of the variance using principal component analysis. We apply the pretrained 18-layer ResNet on the aligned face images and take the output of its last average pooling layer as the high-level visual features. Regarding acoustic features, an acoustic baseline feature set consisting of 88 features is used in this work. The acoustic features consist of the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [46] are extracted by openSMILE. On this basis, two arithmetic functions, arithmetic mean and standard derivation are applied over a fixed window of 8 seconds with a step size of 40 ms on continuous Low-Level Descriptors (LLDs) such as mel-frequency cepstral coefficients, pitch, energy and loudness. Furthermore, referring to [53], the pretrained VGGish model is employed to extract high-level spectral features from raw audio data. Subsequently, these features are integrated with aforementioned handcrafted features to establish the unified representation of the auditory modality. In summary, three feature streams comprising high-level visual features, geometric features, and unified acoustic features extracted from both visual and aural modalities are employed (Table 1).

For the experiments carried out on AVEC2019 dataset, we adopt 69-dimensional eGeMAPS features and 53-dimensional handcrafted visual features provided by this challenge. Handcrafted visual features consist of the descriptors of pose and gaze, the intensities of 17 Facial Action Units (FAUs) along with a confidence measure. Furthermore, the pretrained 18-layer ResNet is applied on the aligned face images to extract the high-level visual features at the frame rate of 50 frames per second. Then the mean features of every 5 frames is adopted as the input. For the acoustic modality, the pretrained VGGish model is also used to extract 128-dimensional high-level spectral features.

4.2 Evaluation metric

In this paper, the Concordance Correlation Coefficient (CCC) metric officially adopted by AVEC2016 is used to evaluate the performance of our proposed model. Compared to Pearson's Correlation Coefficient (PCC), CCC measures the agreement between two sequences while also taking into account numerical precision, both of which are particularly relevant for evaluating the performance of continuous emotion recognition model. CCC is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_y - \mu_x)^2} \tag{11}$$

where ρ denotes the PCC value between the two time series x and y; x represents the true emotion labels' time series, y represents the predicted emotion labels' time series; μ and σ represents the mean and standard deviation of the corresponding time series, respectively. The range of the CCC

Table 1 Summary of dimensional emotion datasets adopted in the evaluation

Dataset	Modalities	Number of subjects	Number of frames per subjects	Annotation	Dimensions
RECOLA [33]	video, audio, physiology	18	7501	frame-level(25 HZ)	arousal, valence
SEWA [34]	video, audio	48	467-1757	frame-level(10 HZ)	arousal, valence, liking

metric is between $[-1, +1]$, with larger values indicating stronger correlation, which -1 means complete inconsistency, 0 no consistency at all, and $+1$ complete consistency.

4.3 Implementation details

Our proposed model is implemented via PyTorch and trained on the Nvidia Geforce RTX 2080Ti Graphics Processing Units (GPUs). This subsection details the important experimental setup. During the training phase, random rotations from -10° to 10°, horizontal flipping, cropping and scaling on aligned face images were performed to avoid severe overfitting. During the training and testing phases, normalization is applied to the facial images, geometric features and unified acoustic features. To extract the aligned face images from video frames, the following three steps are performed. Firstly, OpenCV is adopted to frame each video sample in the dataset. Secondly, due to improper operations in the process of data collection or video encoding, the number of video frames in some video samples is less than the expected number at the video frame rate. Therefore, we complement the naturally missing frames after framing. Finally, referring to [44], aligned face images from video frames are extracted by using OpenFace for face detection and landmark alignment. Due to the failure of face detection in video frames, the corresponding data is filtered out to avoid severe misalignment of sequential data from different modalities. Referring to [54], the annotation delay between recordings and labels in the dataset can seriously damage the emotion recognition system. When predicting valence and arousal, for RECOLA, the delay time is set to 3 seconds and 2.4 seconds to further align feature sequences and labels, respectively. For SEWA, the delay time is set to 2.4 seconds and 1.4 seconds for valence and arousal.

Considering the video length and computational efficiency, the length of the input sequence is set to 500 time steps for RECOLA and 300 time steps for SEWA. For the local contextual temporal convolutional network, we stack 4 layers of local temporal convolution blocks with the kernel size of 3 and set dilation rates of the blocks to $\{1, 2, 4, 8\}$. For RECOLA, the global contextual temporal convolutional network is stacked by 6 layers of global temporal convolution blocks with the kernel size of 7 and dilation rates of the blocks are $\{1, 2, 4, 8, 16, 32\}$. For SEWA, the global contextual temporal convolutional network is stacked by 5 layers of global temporal convolution blocks with the kernel size of 5 and dilation rates of the blocks are $\{1, 2, 4, 8, 16\}$. And the input and output channels in each layer of above networks are set to 128 for arousal and 64 for valence. In the experiments, two different models are independently trained for the prediction of arousal and valence dimensions. To achieve the best performance, we used Adam as the optimizer to train these models and tested the initial learning rate of 0.0001, 0.001 and 0.01 for valence and arousal. The batch size is set to 8 and 16 for RECOLA and SEWA respectively.

4.4 Results and discussion

4.4.1 Comparison with the state-of-the-art works

Table 3 summarizes the results of our proposed approach compared to the reported state-of-the-art works on the original validation set of AVEC2016. This is because the labels of the original test set of AVEC2016 are not released publicly. For the RECOLA dataset, the sizes of trainable parameters for arousal and valence in the proposed model are 3.7 million and 0.9 million, respectively. During the training phase, the average running time of each epoch of the proposed model on the RECOLA dataset for arousal and valence was 65.5s and 48.2s, respectively. The average execution time of our model framework at the testing phase on the RECOLA dataset ranges from 1.3 seconds to 2.4 seconds per iteration. As can be seen from Table 2, our proposed method achieves better performance. In [55], the authors proposed a model framework combining improved AlexNet network and attention mechanism for audio-visual dual-modal emotion recognition. The model utilized the multimodal attention mechanism to fuse hand-extracted audio features and visual features extracted by the AlexNet network. Our method outperforms this fusion method by an average of 2.9%. The method proposed in [56] concatenated visual and audio features learned through a self-supervised strategy into a latent representation vector, and fed it into the bidirectional GRU network to predict arousal and valence values. The CCC val-

Table 2 Multi-modal emotion recognition results and comparison in terms of CCC with the state-of-the-art works on the RECOLA (AVEC 2016) dataset

Method	Fusion level	Arousal (CCC)	Valence (CCC)	Average (CCC)
MAF-LSTM [55]	feature-level	0.729	0.718	0.724
Concat-BGRU [56]	feature-level	0.770	0.464	0.617
EA-BAF [57]	decision-level	0.620	0.720	0.670
AWN-LSTM [32]	model-level	0.830	0.623	0.727
Distilled-LSTM [33]	model-level	0.810	0.630	0.720
Our proposed	model-level	**0.834**	**0.671**	**0.753**

Table 3 Multi-modal emotion recognition results and comparison in terms of CCC with the state-of-the-art works on the Hungarian culture of SEWA (AVEC2019) dataset

Method	Fusion level	Arousal (CCC)	Valence (CCC)	Average (CCC)
ADA-LSTM [58]	feature-level	0.585	0.463	0.524
LSTM-DNN [38]	feature-level	0.513	0.588	0.551
CEDF-LSTM [31]	decision-level	0.534	0.572	0.553
MulT-LSTM [39]	model-level	0.530	0.549	0.540
Our proposed	model-level	0.573	0.533	0.553

ues produced by this method are significantly lower than our results in both the arousal dimension and the valence dimension. In [57], audio and visual features extracted by the embedding attention were separately fed into the LSTM to get the corresponding initial predictions, which were then fed into the proposed decision fusion method to obtain the final results. It provides an average CCC value that is 8.3% lower than the average CCC value generated by our method. In [32], the authors utilized a model-level fusion method based on the adaptive weight network to fuse Local Gabor Binary Pattern-Three Orthogonal Planes (LGBP-TOP) features, Local Phase Quantization-Three Orthogonal Planes (LPQ-TOP) features, geometric features and acoustic features from audio-visual modalities. LSTM model was also adopted in their work to capture long-range contextual information. Their reported CCC values are on average 2.6% lower than those produced by our method. The recognition performance of our proposed method also outperforms the multi-modal continuous emotion recognition method based on LSTM in [33]. Compared to [55–57], we learn importance weights for each input feature stream by using intra-modal attention to highlight effective emotional features, while building the inter-modal attention module to take full advantage of complementary emotional information from multiple modalities. Compared to [32, 33], in addition to effectively fusing emotional features from different modalities, our proposed method also utilizes temporal convolutional networks to capture temporal context information at different scales, which is beneficial for improving the performance of continuous emotion recognition.

Table 3 summarizes the results of our proposed approach compared to the reported state-of-the-art works on the original validation set of AVEC2019. For the SEWA dataset, the sizes of trainable parameters for arousal and valence in the proposed model are 3.1 million and 0.8 million, respectively. And the average execution time of our model framework at the testing phase on the SEWA dataset ranges from 1.2 seconds to 1.9 seconds per iteration. Deng et al. [31] first combined unimodal features into a multi-modal feature vector to train a DBLSTM, and then fed the outputs of multiple unimodal and multi-modal DBLSTMs into a subsequent DBLSTM model to obtain the final prediction results. Chen et al. [38] used concatenation operation to fuse multiple feature vectors and employed LSTM to learn temporal contextual representation in the fused feature sequence. Then, a DNN-based regressor was adopted to predict arousal or valence dimension. Compared to the above approaches, our proposed method achieves comparable recognition results. Huang et al. [39] adopted the multi-head attention mechanism to fuse complementary emotional information between audio-visual modalities in model-level fusion, and further combined transformer and LSTM to explore high-level emotional representation. Their reported CCC values are on average 1.3% lower than those produced by our method. Zhao et al. [58] utilized several pre-trained models to extract efficient deep learning features from acoustic, visual and textual modalities, which were concatenated and fed into LSTM to obtain prediction results in different emotional dimensions. The average CCC value provided by this approach is 2.9% lower than the average CCC value generated by our method.

4.4.2 Detailed analysis

In this subsection, we first conduct ablation experiments on the RECOLA dataset to demonstrate the effectiveness of the proposed model-level fusion method and temporal modeling approach. The temporal convolutional network [43] with non-causal convolution is adopted as the baseline to obtain temporal contextual information, and the concatenation-based feature-level fusion method is adopted as the baseline to fuse the three input feature streams. From Table 4, it can be seen that our multi-modal attention fusion method achieves

Table 4 Ablation study of the proposed Multi-modal Attention Fusion module (MAF) in terms of CCC on RECOLA (AVEC 2016) dataset

Method	Arousal (CCC)	Valence (CCC)	Average (CCC)
Concat(baseline)	0.739	0.571	0.655
MAF w/o MIM	0.826	0.659	0.743
MAF	**0.834**	**0.671**	**0.753**

Table 5 Ablation study of the proposed Local/Global Contextual TCN (LC/GC-TCN) in terms of CCC on RECOLA (AVEC 2016) dataset

Method	Arousal (CCC)	Valence (CCC)	Average (CCC)
TCN(baseline)	0.805	0.623	0.714
LC/GC-TCN w/o Dense	0.822	0.661	0.742
LC/GC-TCN	**0.834**	**0.671**	**0.753**

a significant improvement in the prediction performance of arousal and valence compared to feature concatenation, which demonstrates the advantages of the proposed approach in fusing complementary emotional information from different modalities. Furthermore, the effect of introducing the Modal Interaction Matrix (MIM) or not on the recognition performance is investigated. The results show that by further applying the modal interaction matrix, the CCC values of the arousal and valence dimensions are improved by 0.8% and 1.2%, respectively. This is mainly because it can significantly enhance the ability of our fusion method to model the dynamic interaction of information between modalities. Meanwhile, as shown in Table 5, it can be noticed that utilizing multi-scale temporal information in temporal convolutional networks improves the recognition performance of dimensional affective states. Furthermore, Table 5 illustrates the prediction accuracy of applying dense connections (Dense) to the global context temporal convolutional network. This structure brings performance improvements of 1.2% and 1% for arousal and valence, respectively. This is mainly due to its ability to continuously integrate multi-scale temporal contextual information from different layers for predicting affective states.

Additionally, we evaluate the proposed model framework by using the unimodal feature stream or the combination of multiple feature streams as its input to explore the contribution of different input feature streams to the predicted valence and arousal values. In order to accommodate the experimental requirements, a single-modal emotion recognition architecture is constructed by removing the multi-modal attention fusion module in the multi-modal recognition framework. Table 6 shows the recognition performance obtained by the proposed model when using unimodal and multi-modal data as input. Several conclusions can be drawn from Table 6. Firstly, unified audio features have better recognition performance than geometric features and high-level visual features in predicting arousal values, while geometric features have the best recognition performance in predicting valence values compared to the other unimodal features. Therefore, this result once again proves that the emotional information contained in multiple modalities is complementary and redundant. Secondly, the proposed model framework for emotion recognition using multi-modal data outperforms the unimodal emotion recognition architecture, again illustrating that the proposed multi-modal attention fusion module facilitates the integration of emotional information from different modalities. Thirdly, the best recognition results are achieved in both arousal and valence when using the three feature streams as input, which validates the effectiveness of the proposed overall framework.

To more intuitively illustrate the effectiveness of our proposed method, we present the predicted arousal values for several samples in Fig. 6. It can be seen that the CCC values generated by our proposed method for video 2 and video 6 are as high as 0.856 and 0.881, respectively. Similarly, the valence values predicted by the proposed method for several samples are also listed in Fig. 7. It can be seen that the CCC values of our proposed method are as high as 0.779 and 0.731 for video 1 and video 6, respectively. These all fully demonstrate that the proposed method has good accuracy in predicting continuous affective states.

The framework in research provides a novel direction for improving emotion recognition systems by integrating the the concept of temporal multi-scale into context dependent mod-

Table 6 Recognition performance of unimodal and multi-modal input streams in terms of CCC on RECOLA (AVEC2016) dataset

Features			Method	Arousal (CCC)	Valence (CCC)	Average (CCC)
Audio Unified	Video Geometry	Video CNN-based				
✓			Single	0.729	0.303	0.516
	✓		Single	0.367	0.531	0.449
		✓	Single	0.296	0.436	0.366
✓	✓		Multi-Modal	0.809	0.615	0.712
✓		✓	Multi-Modal	0.801	0.514	0.658
	✓	✓	Multi-Modal	0.506	0.591	0.549
✓	✓	✓	Multi-Modal	**0.834**	**0.671**	**0.753**

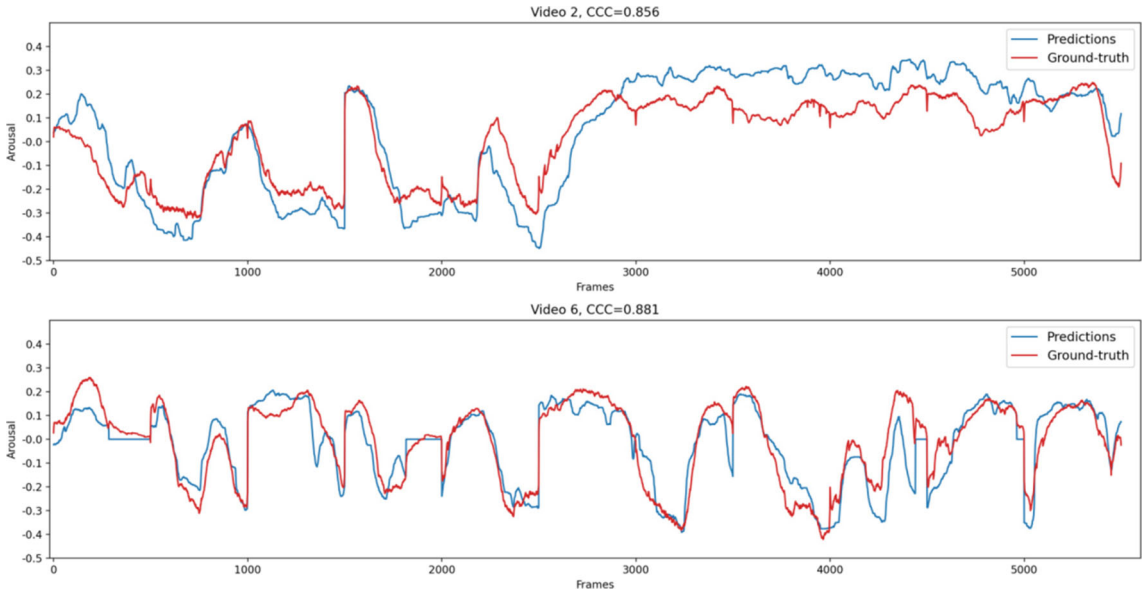

Fig. 6 Visualization of arousal predictions and comparison with the Ground-truth of random samples on RECOLA (AVEC 2016)

eling. Nevertheless, there is room for further improvement in our temporal modeling approach, as the LC/GC-TCN network integrates multi-scale context information at a relatively coarse granularity without fine-grained filtering. Additionally, the model currently treats the prediction of arousal and valence dimensions as independent regression problems, even though these dimensions are intrinsically interrelated. The model framework can be further improved to more effectively integrate multi-scale contextual information and explore the interdependence between different emotional dimensions. Furthermore, the inclusion of the uncertainty for clean datasets is high. The concern of the uncertainty of measured method then is totally missed. Future work can conduct uncertainty analysis on models, which can improve their credibility and application value [59].

5 Conclusion

Continuous emotion recognition remains a challenging task due to the complexities of effectively fusing emotional information from multiple modalities and modeling contextual

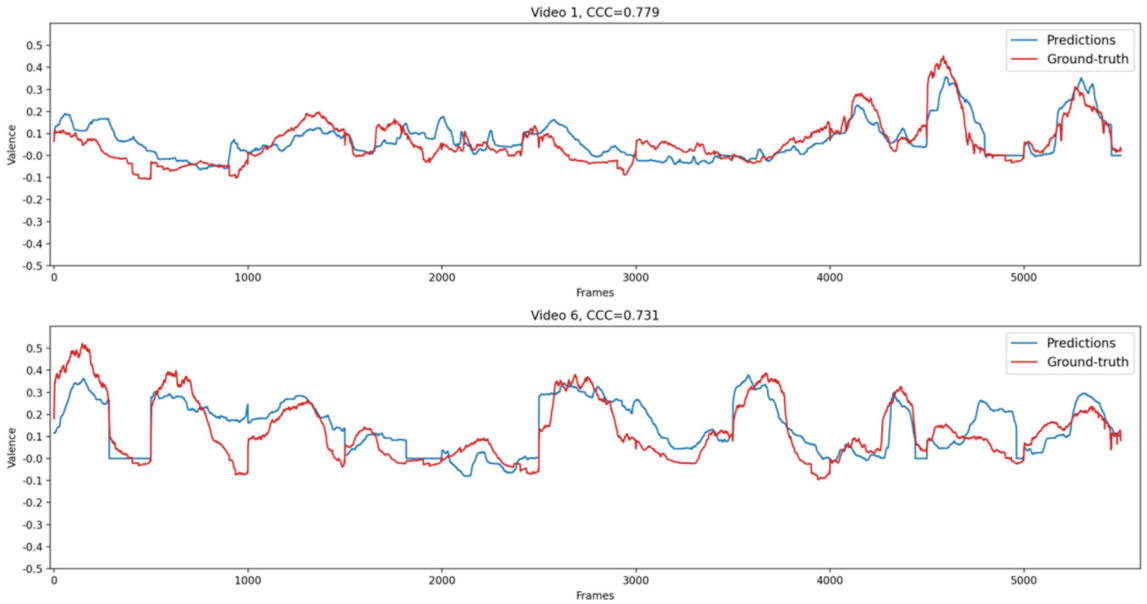

Fig. 7 Visualization of valence predictions and comparison with the Ground-truth of random samples on RECOLA (AVEC 2016)

dependencies during emotional evolution. These challenges have impeded the wider application of the research area's results in human-computer interaction systems. In this paper, an innovative and effective model framework is introduced to address these challenges. Considering the nature that affective states evolve over time and exist multiple representations, the framework in research delves into complex interaction relationships and long-range contextual dependencies within audio-visual modalities. For multi-modal fusion, the multi-modal attention fusion module is proposed to fuse complementary emotional information from different modalities. Specifically, the intra-modal attention is employed to assess the importance of different features within each input feature stream for continuous emotion recognition, thereby highlighting salient emotional features. Subsequently, the inter-modal attention module learns complex correlations among different modalities and facilitates dynamic interactions between them. In terms of temporal modeling, local and global temporal convolutional networks (LC-TCN and GC-TCN) are introduced by stacking temporal convolution blocks. These networks progressively learn ultra-long-range dependencies and capture contextual information at different temporal scales. Consequently, our model establishes complicated mapping relationships between multiple input feature streams and dimensional emotion states. Extensive experiments are conducted on RECOLA and SEWA datasets to show the effectiveness of our proposed model. Compared with the reported state-of-the-art methods, approach in our research achieves better recognition performance. Corresponding results also demonstrate the ability of the proposed method to effectively fuse information from different modalities and capture multi-scale temporal context dependencies for multi-modal continuous emotion recognition.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.U2133218), the National Key Research and Development Program of China (No.2018YFB0204304) and the Fundamental Research Funds for the Central Universities of China (No.FRF-MP-19-007 and No. FRF-TP-20-065A1Z).

Data Availability The data that support the findings of this study are available from https://diuf.unifr.ch/main/diva/recola/ and https://db.sewaproject.eu/ but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of corresponding authors of the RECOLA dataset and the SEWA dataset.

Declarations

Conflict of Interest We declare that we have no actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations that can inappropriately influence our work.

References

1. Bhosale YH, Patnaik KS (2023) Puldi-covid: Chronic obstructive pulmonary (lung) diseases with covid-19 classification using ensemble deep convolutional neural network from chest x-ray images to minimize severity and mortality rates. Biomed Signal Process Control 81(104):445. https://doi.org/10.1016/j.bspc.2022.104445
2. Zhang J, Feng W, Yuan T et al (2022) Scstcf: spatial-channel selection and temporal regularized correlation filters for visual tracking. Appl Soft Comput 118(108):485. https://doi.org/10.1016/j.asoc.2022.108485
3. Zepf S, Hernandez J, Schmitt A et al (2020) Driver emotion recognition for intelligent vehicles: A survey. ACM Computing Surveys (CSUR) 53(3):1–30. https://doi.org/10.1145/3388790
4. Fei Z, Yang E, Li DDU et al (2020) Deep convolution network based emotion analysis towards mental health care. Neurocomputing 388:212–227. https://doi.org/10.1016/j.neucom.2020.01.034
5. Wang W, Xu K, Niu H et al (2020) Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. Complexity 2020:1–9. https://doi.org/10.1155/2020/4065207
6. Zhang J, Yin Z, Chen P et al (2020) Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion 59:103–126. https://doi.org/10.1016/j.inffus.2020.01.011
7. Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76. https://doi.org/10.1016/j.specom.2019.12.001
8. Jiang Y, Li W, Hossain MS et al (2020) A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Information Fusion 53:209–221. https://doi.org/10.1016/j.inffus.2019.06.019
9. Li X, Lu G, Yan J et al (2022) A multi-scale multi-task learning model for continuous dimensional emotion recognition from audio. Electronics 11(3):417. https://doi.org/10.3390/electronics11030417
10. Kollias D, Zafeiriou S (2020) Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. IEEE Trans Affect Comput 12(3):595–606. https://doi.org/10.1109/TAFFC.2020.3014171
11. Rouast PV, Adam MT, Chiong R (2019) Deep learning for human affect recognition: Insights and new developments. IEEE Trans Affect Comput 12(2):524–543. https://doi.org/10.1109/TAFFC.2018.2890471
12. Wang Y, Song W, Tao W et al (2022) A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion. https://doi.org/10.1016/j.inffus.2022.03.009
13. Zhao J, Li R, Chen S et al (2018) Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In: Proceedings of the 2018 on audio/visual emotion challenge and workshop, pp 65–72. https://doi.org/10.1145/3266302.3266313
14. Hao M, Cao WH, Liu ZT et al (2020) Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing 391:42–51. https://doi.org/10.1016/j.neucom.2020.01.048
15. Li C, Bao Z, Li L et al (2020) Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. Inform Process & Manag 57(3):102,185. https://doi.org/10.1016/j.ipm.2019.102185
16. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008

17. Jiang J, Chen Z, Lin H et al (2020) Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: Proceedings of the AAAI conference on artificial intelligence, pp 11,101–11,108, https://doi.org/10.1609/aaai.v34i07.6766
18. Lee J, Kim S, Kim S et al (2020) Multi-modal recurrent attention networks for facial expression recognition. IEEE Trans Image Process 29:6977–6991. https://doi.org/10.1109/TIP.2020.2996086
19. Chen Y, Liu L, Phonevilay V et al (2021) Image super-resolution reconstruction based on feature map attention mechanism. Appl Intell 51:4367–4380. https://doi.org/10.1007/s10489-020-02116-1
20. Antoniadis P, Pikoulis I, Filntisis PP et al (2021) An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3645–3651. https://doi.org/10.1109/ICCVW54120.2021.00407
21. Peng Z, Dang J, Unoki M et al (2021) Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech. Neural Netw 140:261–273. https://doi.org/10.1016/j.neunet.2021.03.027
22. Lee J, Kim S, Kiim S et al (2018) Spatiotemporal attention based deep neural networks for emotion recognition. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 1513–1517. https://doi.org/10.1109/ICASSP.2018.8461920
23. Liu S, Wang X, Zhao L et al (2021) 3dcann: A spatio-temporal convolution attention neural network for eeg emotion recognition. IEEE J Biomed Health Inform 26(11):5321–5331. https://doi.org/10.1109/JBHI.2021.3083525
24. Farha YA, Gall J (2019) Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3575–3584. https://doi.org/10.1109/CVPR.2019.00369
25. Hu M, Chu Q, Wang X et al (2021) A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. IEEE Signal Process Lett 28:698–702. https://doi.org/10.1109/LSP.2021.3063609
26. McKeown G, Valstar M, Cowie R et al (2011) The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput 3(1):5–17. https://doi.org/10.1109/T-AFFC.2011.20
27. Ringeval F, Sonderegger A, Sauer J et al (2013) Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, pp 1–8. https://doi.org/10.1109/FG.2013.6553805
28. Kossaifi J, Walecki R, Panagakis Y et al (2019) Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. IEEE Trans Pattern Anal Mach Intell 43(3):1022–1040. https://doi.org/10.1109/TPAMI.2019.2944808
29. Huang Z, Dang T, Cummins N et al (2015) An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp 41–48. https://doi.org/10.1145/2808196.2811640
30. Nguyen D, Nguyen DT, Zeng R et al (2021) Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. IEEE Trans Multimedia 24:1313–1324. https://doi.org/10.1109/TMM.2021.3063612
31. Chen H, Deng Y, Cheng S et al (2019) Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop, pp 19–26. https://doi.org/10.1145/3347320.3357690
32. Pei E, Jiang D, Sahli H (2020) An efficient model-level fusion approach for continuous affect recognition from audiovisual signals. Neurocomputing 376:42–53. https://doi.org/10.1016/j.neucom.2019.09.037
33. Schoneveld L, Othmani A, Abdelkawy H (2021) Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recogn Lett 146:1–7. https://doi.org/10.1016/j.patrec.2021.03.007
34. Mao Q, Zhu Q, Rao Q et al (2019) Learning hierarchical emotion context for continuous dimensional emotion recognition from video sequences. IEEE Access 7:62,894–62,903. https://doi.org/10.1109/ACCESS.2019.2916211
35. Deng D, Chen Z, Zhou Y et al (2020) Mimamo net: Integrating micro-and macro-motion for video emotion recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 2621–2628
36. Singh R, Saurav S, Kumar T et al (2023) Facial expression recognition in videos using hybrid cnn & convlstm. Int J Inform Technol pp 1–12. https://doi.org/10.1007/s41870-023-01183-0
37. Nagrani A, Yang S, Arnab A et al (2021) Attention bottlenecks for multimodal fusion. Adv Neural Inform Process Syst 34:14,200–14,213. https://doi.org/10.48550/arXiv.2107.00135
38. Chen H, Deng Y, Jiang D (2021) Temporal attentive adversarial domain adaption for cross cultural affect recognition. In: Companion publication of the 2021 international conference on multimodal interaction, pp 97–103
39. Huang J, Tao J, Liu B et al (2020) Multimodal transformer fusion for continuous emotion recognition. In: ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 3507–3511, https://doi.org/10.1109/ICASSP40776.2020.9053762
40. Wu S, Du Z, Li W et al (2019) Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In: 2019 International conference on multimodal interaction, pp 40–48, https://doi.org/10.1145/3340555.3353739
41. Tzirakis P, Chen J, Zafeiriou S et al (2021) End-to-end multimodal affect recognition in real-world environments. Information Fusion 68:46–53. https://doi.org/10.1016/j.inffus.2020.10.011
42. Praveen RG, de Melo WC, Ullah N et al (2022) A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2486–2495, https://doi.org/10.48550/arXiv.2203.14779
43. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In: International conference on learning representations-workshop
44. Du Z, Wu S, Huang D et al (2019) Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE Trans Affect Comput 12(3):565–578. https://doi.org/10.1109/TAFFC.2019.2940224
45. He Z, Zhong Y, Pan J (2022) An adversarial discriminative temporal convolutional network for eeg-based cross-domain emotion recognition. Comput Biol Med 141(105):048. https://doi.org/10.1016/j.compbiomed.2021.105048
46. Eyben F, Scherer KR, Schuller BW et al (2015) The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202. https://doi.org/10.1109/TAFFC.2015.2457417
47. Ruan D, Yan Y, Lai S et al (2021) Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7660–7669
48. Verma S, Wang C, Zhu L et al (2019) Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In: International joint conference on artificial intelligence,

international joint conferences on artificial intelligence organization. https://doi.org/10.24963/ijcai.2019/503
49. Mai S, Xing S, Hu H (2019) Locally confined modality fusion network with a global perspective for multimodal human affective computing. IEEE Trans Multimed 22(1):122–137. https://doi.org/10.1109/TMM.2019.2925966
50. Gao Z, Wang X, Yang Y et al (2020) A channel-fused dense convolutional network for eeg-based emotion recognition. IEEE Trans Cogn Dev Syst 13(4):945–954. https://doi.org/10.1109/TCDS.2020.2976112
51. Ringeval F, Schuller B, Valstar M et al (2019) Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop, pp 3–12. https://doi.org/10.1145/3347320.3357688
52. Valstar M, Gratch J, Schuller B et al (2016) Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th international workshop on audio/visual emotion challenge, pp 3–10. https://doi.org/10.1145/2988257.2988258
53. Zhang S, Ding Y, Wei Z et al (2021) Continuous emotion recognition with audio-visual leader-follower attentive fusion. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3567–3574, https://doi.org/10.48550/arXiv.2107.01175
54. Khorram S, McInnis MG, Provost EM (2019) Jointly aligning and predicting continuous emotion annotations. IEEE Trans Affect Comput 12(4):1069–1083. https://doi.org/10.1109/TAFFC.2019.2917047
55. Liu M, Tang J (2021) Audio and video bimodal emotion recognition in social networks based on improved alexnet network and attention mechanism. J Inform Process Syst 17(4):754–771
56. Shukla A, Petridis S, Pantic M (2023) Does visual self-supervision improve learning of speech representations for emotion recognition. IEEE Trans Affect Comput 14(1):406–420. https://doi.org/10.1109/TAFFC.2021.3062406
57. Lucas J, Ghaleb E, Asteriadis S (2020) Deep, dimensional and multimodal emotion recognition using attention mechanisms. In: BNAIC/BeneLearn 2020, pp 130
58. Zhao J, Li R, Liang J et al (2019) Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop, pp 37–45. https://doi.org/10.1145/3347320.3357692
59. Abbaszadeh Shahri A, Shan C, Larsson S (2022) A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning. Nat Resour Res 31(3):1351–1373. https://doi.org/10.1007/s11053-022-10051-w

Congbao Shi is currently pursuing a master's degree at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. His research interests focus on multi-modal learning, emotional computing.

Yuanyuan Zhang is currently a postdoc at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. Her research interests focus on emotional computing.

Baolin Liu Ph.D., Professor of University of Science and Technology Beijing. Vice President of Zhongguancun Industrial Research Institute of Intelligent Science and Technology. Vice Director of Beijing International Science and Technology Cooperation Base for Cybermatics and Cyberspace. His research interests focus on cognitive computing, emotional computing, machine learning and intelligent information processing.