Multimedia Tools and Applications (2023) 82:7319–7341
https://doi.org/10.1007/s11042-022-13577-6

Modularized composite attention network for continuous music emotion recognition

Meixian Zhang [1] · Yonghua Zhu [1] · Wenjun Zhang [1,2] · Yunwen Zhu [1] · Tianyu Feng [1]

Received: 10 February 2021 / Revised: 15 June 2022 / Accepted: 18 July 2022 /
Published online: 19 August 2022

Abstract

Music Emotion Recognition (MER) has attracted much interest in the past decades. Many deep learning methods have been applied to this field recently. However, the previous methods for MER mostly utilized simple convolutional layers to extract features from the original audio signals, in which representative emotion-related features cannot be extracted. In this paper, we propose a novel method named Modularized Composite Attention Network (MCAN) for continuous MER. A sample reconstruction technique is proposed to enhance the stability of the network. Specifically, a feature augmentation module is constructed to extract salient features and we design a weighted attention module to control the focus of the whole network. Furthermore, a style embedding module is introduced to enhance the detail processing capability of the network. We conduct experiments on two datasets, that is, the benchmark dataset DEAM and the newly proposed dataset PMEmo. The superior results prove the effectiveness of our proposed MCAN. Especially qualitative analyses are given to for explaining the performance of our model.

Keywords Music emotion recognition · Filter bank output · Handcrafted features · Valence · Arousal

✉ Yonghua Zhu
zyh@shu.edu.cn

Meixian Zhang
little_xx@shu.edu.cn

Wenjun Zhang
18096@gench.edu.cn

Yunwen Zhu
eilleen31@shu.edu.cn

Tianyu Feng
fengtianyu@shu.edu.cn

[1] Shanghai Film Academy, Shanghai University, Shanghai, China

[2] College of Information Technology, Shanghai Jianqiao University, Shanghai, China

1 Introduction

The demand for services that provide perceptual interaction capabilities has made affective computing much more vital nowadays. Music is highly associated with human life due to the ability to induce and convey emotions [51]. The development of digital technology has greatly contributed to the tremendous growth of digital music libraries. The way that music information is organized and retrieved has to update to satisfy the ever-increasing need for effective information access [48]. As a consequence, the study of MER has become a hotspot and can be utilized to improve Music Information Retrieval (MIR). The study of social tagging on Last.fm[1] shows that emotional tags are the third most frequently used category when people search for music, which also indicates the importance of MER for MIR [48]. Furthermore, emotion in music also plays an important role in music recommendation [13], therapy [11], interaction [20], music automatic generation, and automatic soundtrack [55]. In general, MER can be applied for automatic emotional annotation of music pieces, which can provide technical support for different application scenarios, such as music information recommendation, cross-modal information interaction, and treatment of symptoms such as depression, etc. From a long-term perspective, music is closely related to the expression of emotions, and MER will become one of the important emotional interaction tools under the environment of intelligent human-computer interaction.

Many participants have obtained considerable results on Music Emotion Classification (MEC) task in the annual campaign at Music Information Retrieval Evaluation eXchange (MIREX)[2] since 2007. In the earlier time, researchers utilized machine learning methods in MER, in which handcrafted features were needed. The difficulties of traditional algorithms, such as Supported Vector Regression (SVR) [31] and Gaussian Mixture Model (GMM) [43], are the extraction and selection of handcrafted features. However, the extraction process requires a lot of music prior knowledge [10]. The omission of necessary features and the inclusion of irrelevant features will both lead to undesirable results. Since Convolutional Neural Networks (CNNs) could be applied to extract features from the original picture on image classification tasks, researchers started to employ CNNs to automatically learn features in audio processing tasks. The studies of Orjesek et al. [30] and Mao et al. [28] suggested that CNNs similarly were appropriate to extract features from spectrograms but those features outperformed handcrafted ones. Sarkar et al. [34] utilized VGGNet with reduction of layers for music emotion classification task. This method has achieved better performance than traditional algorithms. However, methods based on CNNs ignored the sequential information, for which the contextual relationship cannot be used effectively. RNN due to its ability to process sequential data has shown great advantages in MER. Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) have widely been used in MER. Referring to the results of the "Emotion in Music" task at MediaEval [1], methods with LSTM modules have achieved state-of-the-art performance but still had difficulties in the extraction of handcrafted features. Then, researchers attempted to combine the automatic feature extraction of CNNs and the sequential information processing of RNN. A fusion model proposed by Dong et al. [12] verified the effectiveness of combining CNNs and RNN. However, the existing fusion methods only adopted simple CNNs and RNN, which cannot extract salient features and did not take the deep relationship of sequential series into consideration.

[1] Last.fm. Available: https://www.last.fm/
[2] MIREX. Available: http://www.music-ir.org/mirex/wiki/

According to Yang et al. [48], MER can be divided into two types. One is the MEC task that emotions of music are sorted into specific classes. The other is the music emotion regression task that annotations of music pieces are numerical values in two-dimensional (valence and arousal) or three-dimensional (valence, arousal, and dominance) space. The obtain of ground-truth for music emotion is more subjective than other annotations (e.g., genre, instrument). Dividing music emotion into numbers of classes is insufficient due to the variety of emotion categories and the subjectivity of annotations. Continuous annotations can realize the sufficient representation of music emotion by annotating the music pieces in numerical ways [51]. Furthermore, referred to MacDorman et al. [26], the consistency of annotations was low because the added dimension would lead to heavy cognitive burden on annotators, so the expansion in dimension cannot prompt the recognition accuracy much. Thus, the study of continuous MER in two dimensions is more reasonable.

The continuous MER task is to extract features based on raw audio, and propose a method to learn the mapping relationship between music pieces and numerical emotion labels. In this paper, we propose Modularized Composite Attention Network (MCAN) for continuous MER. We take advantage of the adaptive salient-feature learning of CNNs and the weighting ability of attention mechanism to optimize our model. As is shown in Fig. 1, the emotional model we choose to represent emotions is a fuzzy dimensional model with valence (negative to positive) and arousal (silent to energetic) axis proposed by Jun et al. [21], which is modified from Thayer's two-dimensional emotional model [40]. Inspired by the mixup [52] technique, we propose a sample reconstruction technique to produce the noise input, which can highly enhance the stability of the network. We construct a feature augmentation module based on the two branches of filter bank output and handcrafted features. Attention mechanism is introduced to extract salient features. A style embedding module is introduced to provide sufficient emotion-related details. The Concordance Correlation Coefficient (CCC) is used to construct the loss function to prompt the performance of MCAN.

The main contributions of our work are summarized as follows:

1) We design a sample reconstruction technique. To combat the subjectivity of sentimental labels, we construct noisy samples by random resampling through origin sampling in dataset. The reconstructed samples will be input into the network as noise input. This technique can help to enhance the stability of the network.
2) We propose a feature augmentation module. The filter bank output and handcrafted features are extracted as the inputs of the network. Then we design an early fusion architecture to make the two branches of features integrated together. This module can help to provide more property details of music pieces.
3) We propose a weighted attention module for the extraction of salient features. Instead of simply weighting by attention mechanism, the original feature map is added into the weighted one to get the balanced feature map. This module can help to control the focus of the network.
4) We construct a style embedding module. The metadata of music piece is transformed into vector and fed into the regression part of the network. This module can help to strengthen the ability of processing details of network.

The remaining of this paper is organized as follows: Section 2 demonstrates the related works of MER focusing on different targets implemented by various algorithms. Section 3 describes

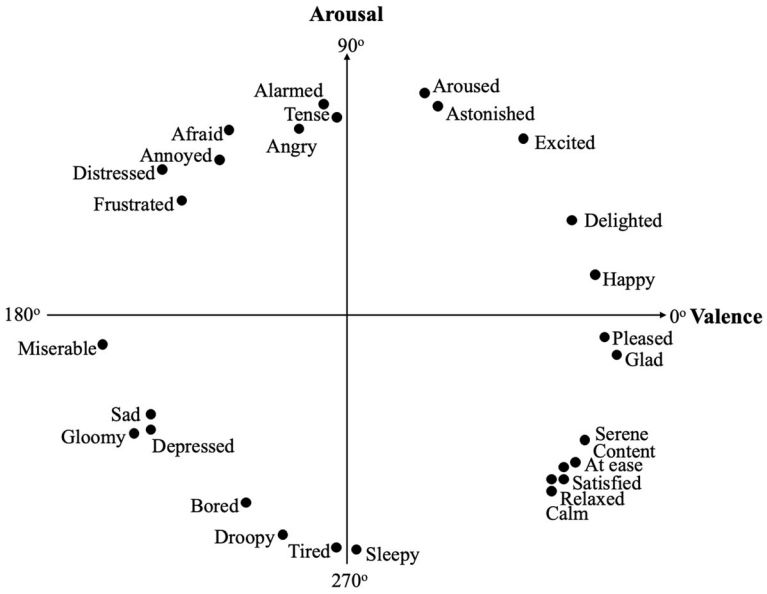

Fig. 1 The valence-arousal emotion model

the details of MCAN. Section 4 represents the specific configurations of our experiments and analysis. Finally, we conclude our work and summarize the future work in Section 5.

2 Related work

Over the past decades, although it is a newly emerging field, the research in MER has developed rapidly. Many machine learning methods have been applied to MER with the extraction and selection of handcrafted features in the earlier time. Researchers gradually adopted algorithms that performed well in image processing to audio processing. Consequently, many architectures based on CNNs have been widely used in MER. As is mentioned above, MER can be classified into two types: categorical recognition and numerical prediction. In the following paragraphs of this section, we will introduce the details of different ways for feature extraction and methods for recognition applied in MER.

2.1 Feature extraction

Different emotional states of music are tightly associated with different patterns of acoustic cues [48, 51]. Handcrafted features are essential while applying traditional machine learning algorithms or methods based on RNN in MER. Yang et al. [48] comprehensively summarized the commonly used tools (e.g., PsySound [6], MIRtoolbox [22], Marsyas [41]) for feature extraction and selected feature sets. Researchers developed several libraries based on different programming languages to extract features in five perceptual dimensions of music listening, that is, energy, rhythm, temporal, spectrum, and melody. Grekow et al. [14] introduced the newly proposed feature extraction tool called Essentia [4]. They concluded that features extracted from raw audio were consisted of low-level, middle-level and high-level

representations. In terms of the selection of handcrafted features, researchers usually adopted Principal Component Analysis (PCA) and Factor Analysis (FA) to reduce the dimensionality of features [14, 48]. However, it is difficult to select comprehensive emotion-related features in music only by traditional ways. The reduction of dimensionality can also easily lead to the omission of essential features.

Consequently, many studies have attempted to adopt CNNs to automatically learn features. Since CNNs have presented extraordinary ability for feature extraction in image classification tasks, Mao et al. [28] applied it on spectrograms in speech emotion recognition task. All of these demonstrated that CNNs could be adapted for adaptive feature extraction through raw audio. Dong et al. [12] employed simple CNNs for feature extraction and the results suggested that the extracted features outperformed several handcrafted-feature sets. However, some methods with deep architectures, such as DenseNet [17] and ResNet [15], which saliently performed better than simple CNNs [36] in image recognition, cannot contribute to better results in audio processing tasks due to the insufficient data representation.

Although CNNs have been successfully utilized for feature extraction in MER, previous methods did not take sufficient information of spectrograms into account. To solve these problems, we propose a novel feature extraction module based on the feature augmentation module and the weighted attention module, which can help to extract salient features through sufficient details.

2.2 Music emotion recognition

In categorical approach, emotions in music are annotated by different classes such as the basic emotions (e.g., happy, angry, sad, and relaxing). Researchers assigned a piece of music with only one class, which formed the study of single-label classification. The studies of [3, 33, 42] applied machine learning methods such as SVR, K-Nearest Neighbor (KNN), and MLR. Deep learning methods have also been utilized in MER. Sarkar et al. [34] employed VGGNet in single-label classification and the results suggested that the deep learning methods were applicable in this field. Many novel methods, such as Vector-Quantized Variational Auto-Encoder (VQ-VAE) [19], RNN based on LSTM [2, 29], CNNs based on self-attention mechanism [38], and ResNet based on frequency aware convolution [5], performed well in detecting emotion and theme in music tracks at MediaEval2019.[3] However, single-label is insufficient for representing emotions in music due to the subjectivity of perceptions. Li et al. [23] divided music into different clusters by SVMs. MER was treated as a multi-label classification task in this method. Based on the same classification approach, Wu et al. [46] proposed a novel hierarchical Bayesian model. Unfortunately, because of the cognitive burden of humans and the ambiguous boundaries between different emotions, the annotations for multi-label classification are difficult to obtain. Thus, fuzzy classification is proposed to balance the simplicity of single-label classification and the difficulty in obtaining labels of multi-label classification. Yang et al. [49] firstly attempted to utilize the fuzzy classification method for MER to deal with the subjective issue of annotations. In this approach, the fuzzy vectors combined by the probabilities of different emotions were computed to represent the emotions contained in a piece of music.

The fuzzy classification methods can deal with the subjective issue of emotional expression to a certain extend. However, emotions are insufficient to be expressed with only several classes. There will be a fine-grained issue while treating emotion recognition as a classification task [48]. Thus, many researchers represented emotions in a two-dimensional space as is shown in Fig. 1. They

[3] MediaEval2019. Available: http://www.multimediaeval.org/mediaeval2019/

regarded MER as a numerical prediction task. Yang et al. [50] groundbreakingly applied models such as MLR, SVR, and AdaBoost.RT to numerical regression for MER. There were many state-of-the-art methods based on RNN emerging for continuous prediction tasks since 2015. Orjesek et al. [30] employed GRU as a unit in RNN. Chen et al. [7] proposed a model based on BLSTM for multimodal emotion recognition. Their experimental results suggested that BLSTM performed better than LSTM. The studies of [25, 27] have also shown the superiority of LSTM in processing audio signals. Dong et al. [12] proposed a bidirectional convolutional recurrent sparse network, in which CNN is adopted to extract the feature maps through original spectrograms. In this method, the disadvantages of handcrafted feature extraction could be avoided. Cheuk et al. [8] proposed a novel method in traditional way, which is based on SVR and performed well in regression task. In general, because of the subjectivity of human annotations, assigning one piece of music with only one point in a two-dimensional plane is still not enough. Thus, continuous probability distribution prediction is proposed to match the original annotations. In this approach, the problem of labeling can be solved fundamentally. Schmidt et al. [35] firstly treated MER as an issue of probability distribution. In this method, the probability density function was learned from the original labels. Then, the features extracted from audio and the annotations could be mapped in a linear way. However, models based on probabilistic prediction are more susceptible to the impact of original annotations. Though subjective factors could be eliminated in these methods, the establishment of the dataset would be more rigorous and the application would be more restricted.

From a comprehensive perspective, the dimensional numerical prediction is more applicable and reasonable. Although the previous methods have achieved great improvement in different MER tasks by using deep learning techniques. Some of these methods used simple CNNs to extract features from spectrograms, which ignored semantic relationship. Others did not consider the sequential information. To better extract emotion-related features and employ temporal relationship, we design a method based on the feature augmentation module and the sequential processing module. Furthermore, we design a sample reconstruction technique to enhance the stability of the network and the style embedding module to supply sufficient details.

3 Modularized composite attention network

In this section, we introduce the proposed model MCAN. The architecture of MCAN is shown in Fig. 2, which contains four main parts. The first one is the sample reconstruction technique, which can help to make the network sensitive to slight difference among music pieces with near emotional values. The second one is the feature augmentation module combined by the processing of filter bank output and the handcrafted features. After integrating the two branches of features, the feature map is fed into a weighted module based on self-attention. Then the output will be input into the BLSTM module to process the sequential information. Furthermore, a style embedding module is designed to provide sufficient details related to the expression of emotions. We will introduce the main details of our model in the following paragraphs.

3.1 Sample reconstruction

According to Deshpande et al. [9], a spectrogram contains all the physical information of the original audio. Thus, it can be a sufficient representation of audio signals. Compared to spectrograms, Pons et al. [32] used raw audio as an input, the results of which suggested that raw audio

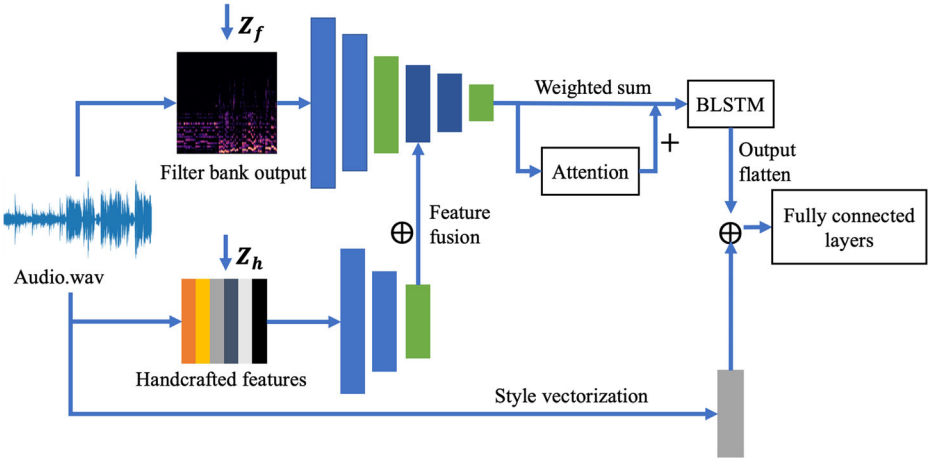

Fig. 2 The whole architecture of the proposed MCAN. $\boldsymbol{Z_f}$ and $\boldsymbol{Z_h}$ are the reconstruction noise calculated through the process of sample reconstruction

cannot greatly improve the performance. Furthermore, the processing of raw audio is more complicated. The filter bank output is an intermediate product between Mel-spectrograms and MFCCs. Compared with Mel-spectrograms, filter bank output includes the log operation, eliminating some redundant information. Compared with MFCCs, filter bank output has not undergone the conversion operation of MFCC and retains more details. Consequently, we extract the filter bank output through raw audio, whose x-axis represents the information of time-domain and y-axis denotes the information of frequency domain, as an input to the networks representing the original characteristics of audio signals. The size of the each extracted filter bank output is 120*120. Furthermore, to solve the problem of insufficient data representation in the existing methods, we extract the handcrafted features for each time window of music piece after windowing, framing, and pre-emphasis. We extract 60 features from the commonly used handcrafted features set [1] by features engineering as our handcrafted feature set. These features are highly associated with the properties of music and can help to strengthen the information acquisition ability of the network. The size of each extracted feature map is 60*60.

Since sentiment annotation is highly subjective, we propose a sample reconstruction method for this problem. The main idea of sample reconstruction is to randomly resample two samples in dataset with different sentiment annotation values, and construct a new sample by linear combination based on these two samples. In addition, we introduce the reconstruction scale parameter α to control the ratio of input noise. The reconstructed samples are input to the network as noise input, and the introduction of noise can make the model more stable. Assume that the sequence of filter bank output is $\boldsymbol{X_F} = [\boldsymbol{x_1}^f, \boldsymbol{x_2}^f, \ldots, \boldsymbol{x_N}^f]$ and the sequence of handcrafted features is $\boldsymbol{X_H} = [\boldsymbol{x_1}^h, \boldsymbol{x_2}^h, \ldots, \boldsymbol{x_N}^h]$, where *f*,*F* denote filter bank output and *h*,*H* denote the handcrafted feature. The sequence of labels is $\boldsymbol{Y} = [y_1{}^v, y_2{}^v, \ldots, y_N{}^v]$ and $\boldsymbol{Y} = [y_1{}^a, y_2{}^a, \ldots, y_N{}^a]$, where *v* denotes the dimension of valence and *a* denotes the dimension of arousal. *N* is the total size of samples. The sample reconstruction of each sample pairs is shown in Fig. 3.

Similar to mixup, the sample reconstruction is generally based on a generic vicinal distribution as follows:

$$\mu\left(\widetilde{\boldsymbol{x}}, \widetilde{y} | \boldsymbol{x_i}, y_i\right) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{E}_{\omega}\left(\mathrm{D}\left(\widetilde{\boldsymbol{x}} = \omega \boldsymbol{x_i} + (1-\omega)\boldsymbol{x_j}, \widetilde{y} = \omega y_i + (1-\omega) y_j\right)\right) \quad (1)$$

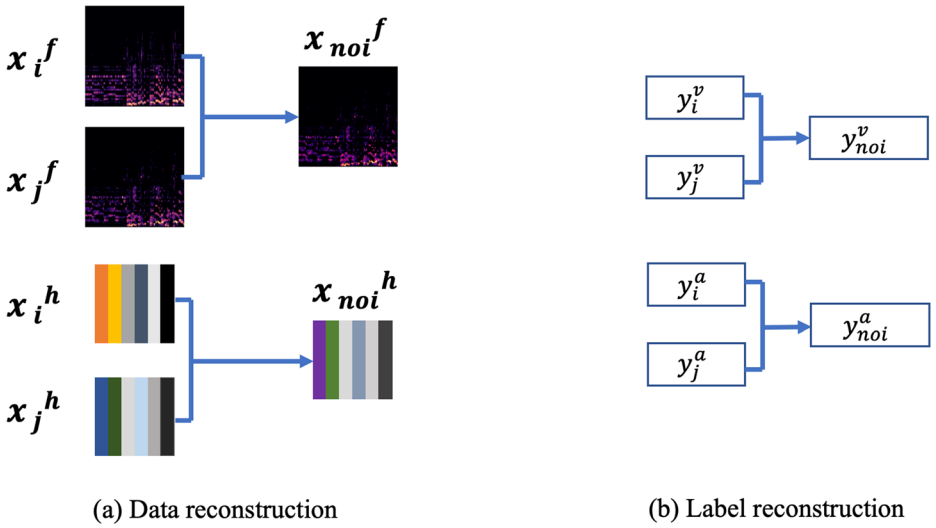

Fig. 3 The sample reconstruction technique applied to produce training noise samples by weighting the original sample pairs by linear interpolation. **a** represents the reconstruction of data pairs and **b** represents the reconstruction of label pairs

where $\mu(\tilde{x},\tilde{y}|\boldsymbol{x_i},y_i)$ denotes the mean value of $\tilde{x},\tilde{y}$ based on interpolation operations for sample pairs ($\boldsymbol{x_i}$, y_i). E_ω represents the expectation calculation for ω. D represents the Dirac mass calculation. ω is calculated by Beta (θ, θ) and $\theta \in (0, +\infty)$. In our sample reconstruction, the noise sample pairs can be produced as follows:

$$\boldsymbol{x}_{\boldsymbol{noi}}^{\boldsymbol{f}} = \omega \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{f}} + (1-\omega)\boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{f}} \tag{2}$$

$$\boldsymbol{x}_{\boldsymbol{noi}}^{\boldsymbol{h}} = \omega \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{h}} + (1-\omega)\boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{h}} \tag{3}$$

$$y_{noi}^{v} = \omega y_i^v + (1-\omega) y_j^v \tag{4}$$

$$y_{noi}^{a} = \omega y_i^a + (1-\omega) y_j^a \tag{5}$$

where $\boldsymbol{x_i}^{\boldsymbol{f}}$, $\boldsymbol{x_j}^{\boldsymbol{f}}$ are the filter bank output pairs, $\boldsymbol{x_i}^{\boldsymbol{h}}$, $\boldsymbol{x_j}^{\boldsymbol{h}}$ are the handcrafted feature pairs, y_i^v, y_j^v are the label pairs in valence dimension, and y_i^a, y_j^a are the label pairs in arousal dimension. $\omega \in [0, 1]$ is a computed weight hyper-parameter. $\boldsymbol{x}_{\boldsymbol{noi}}^{\boldsymbol{f}}, \boldsymbol{x}_{\boldsymbol{noi}}^{\boldsymbol{h}}, y_{noi}^{v}, y_{noi}^{a}$ are the produced noise samples.

After performing sample reconstruction, we have a more robust dataset containing noisy information $\boldsymbol{Z}_f = \left[\boldsymbol{x}_{\boldsymbol{noi}}^{f_1}, \boldsymbol{x}_{\boldsymbol{noi}}^{f_2}, \ldots, \boldsymbol{x}_{\boldsymbol{noi}}^{f_M}\right]$, $\boldsymbol{Z}_h = \left[\boldsymbol{x}_{\boldsymbol{noi}}^{h_1}, \boldsymbol{x}_{\boldsymbol{noi}}^{h_2}, \ldots, \boldsymbol{x}_{\boldsymbol{noi}}^{h_M}\right]$, and $\boldsymbol{Y}_{\boldsymbol{noi}} = \left[\boldsymbol{y}_{\boldsymbol{noi}}^{\boldsymbol{1}}, \boldsymbol{y}_{\boldsymbol{noi}}^{\boldsymbol{2}}, \ldots, \boldsymbol{y}_{\boldsymbol{noi}}^{\boldsymbol{M}}\right]$. $\boldsymbol{Z_f}$ is the noisy data of filter bank output combined by the m-th reconstructed sample $\boldsymbol{x}_{\boldsymbol{noi}}^{\boldsymbol{f}_m}$. M is the total size of noisy samples. $\boldsymbol{Z_h}$ is the similarly reconstructed sample set of handcrafted features. $\boldsymbol{Y_{noi}}$ is the reconstructed noisy label matrix and $\boldsymbol{y}_{\boldsymbol{noi}}^{\boldsymbol{m}} = [y_{noi}^{v_m}, y_{noi}^{a_m}]$ represents the m-th sample labels in valence and arousal dimensions. Finally, the total

size of data is $M + N$. The reconstructed samples can enhance the ability of network to learn the mapping relationship between music pieces with different sentiment values.

3.2 Feature augmentation module

In previous methods, researchers often applied late fusion in feature construction, in which the correlation among information in different perspectives were ignored. To fully exploit these correlations, we design a feature augmentation module. In this module, the processed handcrafted feature maps are embedded into the filter bank output ones, and then they are concatenated together to get the early fusion results. The handcrafted features provide guidance to the feature extraction module in a way of embedding, which enhances the representation of spectrum-based features.

The feature augmentation module is shown in Fig. 4. Given the filter bank output $\boldsymbol{X_F}$ and the handcrafted features $\boldsymbol{X_H}$. The $\boldsymbol{X_F}$ is transformed into feature map $\boldsymbol{X_F}'$ by a two-layer convolutional network and one max pooling layer. The $\boldsymbol{X_H}$ is transformed into feature map $\boldsymbol{X_H}'$ in the similar way. Then the feature map $\boldsymbol{X_F}'$ is reshaped by a convolutional layer to match the size of $\boldsymbol{X_H}'$ and the feature map $\boldsymbol{X_F}''$ is produced. $\boldsymbol{X_F}''$ and $\boldsymbol{X_H}'$ is concatenated together to construct the feature map $\boldsymbol{X_{FH}}$. The augmented feature map $\boldsymbol{X_{FH}}'$ is finally gotten by a combination of convolutional layer and max pooling layer. The whole calculation process is expressed as follows:

$$\boldsymbol{X}'_{\boldsymbol{FH}} = \mathrm{f}_4(\mathrm{f}_2(\mathrm{f}_1(\boldsymbol{X_F}))\oplus \mathrm{f}_3(\boldsymbol{X_H})) \tag{6}$$

where $\mathrm{f}_n(\cdot)$ represents the different calculation in part n. $\oplus$ is the concatenation calculation. The feature augmentation module is based on information in different perspectives and the augmented features are produced by embedding technique, which greatly improves the ability of the network to extract salient features. In this way, the mapping relationship between data and labels can be better learned.

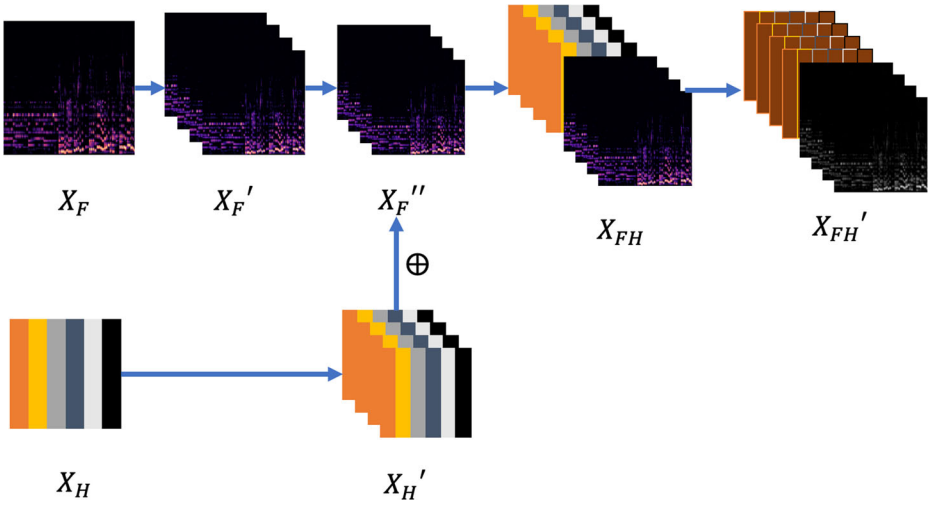

Fig. 4 Feature augmentation module. The process of feature fusion between filter bank output and handcrafted features

3.3 Weighted attention module

Despite the richer details, the lack of focus will increase the cognitive load of the network. To address the problem of cognitive load, we design a weighted attention module based on attention mechanism and BLSTM module. Wang et al. [44] and Huan et al. [16] adopted attention mechanisms in video emotion recognition and verified the ability of them to improve performance. According to the superiority of attention mechanism in specialized data processing, we introduce self-attention mechanism to control the focus of the processing of features. Assume that the input feature map is $\boldsymbol{X}_{\boldsymbol{FH}}^{'} = \left[\boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{1}}, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{2}}, \ldots, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{M+N}}\right]$. A batch of samples at time t in sequence can be expressed as $\boldsymbol{X}_{\boldsymbol{FH}}^{'\,\boldsymbol{t}} = \left[\boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{1}\,\boldsymbol{t}}, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{2}\,\boldsymbol{t}}, \ldots, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{M+N}\,\boldsymbol{t}}\right]$. There are three weight matrixes $\boldsymbol{W}_{\boldsymbol{Q}}$, $\boldsymbol{W}_{\boldsymbol{K}}$, $\boldsymbol{W}_{\boldsymbol{V}}$ as the hyper-parameters learned by the self-attention mechanism to respectively obtain the query, key and value vectors. These vectors are calculated as:

$$q_n^{:t} = W_Q \cdot x_{fh}^{n\,t} \tag{7}$$

$$k_n^{:t} = W_k \cdot x_{fh}^{n\,t} \tag{8}$$

$$v_n^{:t} = W_v \cdot x_{fh}^{n\,t} \tag{9}$$

where $\boldsymbol{q_n}^{:\boldsymbol{t}}$, $\boldsymbol{k_n}^{:\boldsymbol{t}}$, $\boldsymbol{v_n}^{:\boldsymbol{t}}$ are the query, key and value vector in step t and sample n. To better utilize the information carried by the former time steps, the weight vector can be calculated as:

$$a_n^{:t} = \mathrm{softmax}\left(K_n^T, q_n^{:t}\right) \tag{10}$$

where $\boldsymbol{a}_{\boldsymbol{n}}^{:\boldsymbol{t}}$ is the weight vector applied to the value vector to aggregate the sequence information together, for which the contextual information is weighted to analysis the emotion contained in the whole music piece. $\boldsymbol{K}_{\boldsymbol{n}}^{\boldsymbol{T}} = \left[\boldsymbol{k_n}^{:\boldsymbol{1}}, \boldsymbol{k_n}^{:\boldsymbol{2}}, \ldots\ldots, \boldsymbol{k_n}^{:\boldsymbol{T_n}}\right]^T$ denotes the transition of the matrix concatenated by the key vectors and T_n is the length of sequence. T is the transpose operation. Softmax is the activation function. Then, the output cell of the self-attention layer is computed as:

$$\boldsymbol{c}_{\boldsymbol{n}}^{:\boldsymbol{t}} = \sum_{i=1}^{T_n} \boldsymbol{a}_{\boldsymbol{n}}^{\boldsymbol{it}} \cdot \boldsymbol{v}_{\boldsymbol{n}}^{:\boldsymbol{i}} \tag{11}$$

where $\boldsymbol{c}_{\boldsymbol{n}}^{:\boldsymbol{t}}$ is the output of the context-attention layer in time step t and sample n. $\boldsymbol{a}_{\boldsymbol{n}}^{\boldsymbol{it}}$ denotes the i-th element of vector $\boldsymbol{a}_{\boldsymbol{n}}^{:\boldsymbol{t}}$. Then the feature maps processed by self-attention mechanism can be represented as $\boldsymbol{X}_{\boldsymbol{FH}}^{\boldsymbol{attn}} = \left[\boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{attn}_1}, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{attn}_2}, \ldots, \boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{attn}_{M+N}}\right]$. The n-th sample can be represented as $\boldsymbol{x}_{\boldsymbol{fh}}^{\boldsymbol{attn}_n} = \left[\boldsymbol{c}_{\boldsymbol{n}}^{:\boldsymbol{1}}, \boldsymbol{c}_{\boldsymbol{n}}^{:\boldsymbol{2}}, \ldots, \boldsymbol{c}_{\boldsymbol{n}}^{:\boldsymbol{T_n}}\right]$. Then the weighted sum feature maps $\boldsymbol{X}_{\boldsymbol{FH}}^{\boldsymbol{WS}}$ can be calculated as:

$$\boldsymbol{X}_{\boldsymbol{FH}}^{\boldsymbol{WS}} = (1-\beta)\boldsymbol{X}_{\boldsymbol{FH}}^{'} + \beta \boldsymbol{X}_{\boldsymbol{FH}}^{\boldsymbol{atten}} \tag{12}$$

where β is the parameter to control the weight of the original augmented feature maps and the weighted feature maps. The whole calculation process is shown in Fig. 5.

Finally, to better capture the relationship within contextual information, the weighted sum feature maps are input into the BLSTM module for sequential processing. LSTM module, which can solve the issue of gradient vanishing, has a memory cell to store information with an

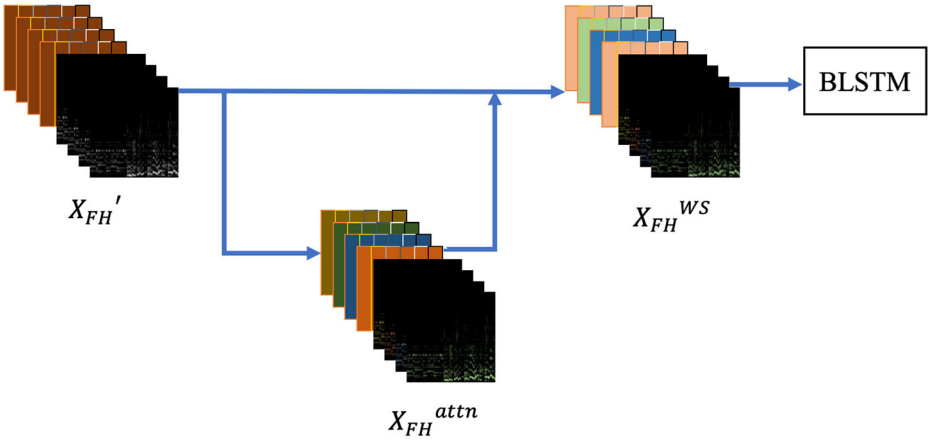

Fig. 5 Weighted attention module. The process of self-attention mechanism, weighted sum and sequential processing

input gate, an output gate and a forget gate. The calculation of the forward BLSTM layer is as the following equations:

$$\boldsymbol{i}_t = \sigma_1(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \tag{13}$$

$$\boldsymbol{o}_t = \sigma_1(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{c}_{t-1} + \boldsymbol{b}_o) \tag{14}$$

$$\boldsymbol{f}_t = \sigma_1(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \tag{15}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \otimes \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \otimes \sigma_2(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{16}$$

$$\boldsymbol{h}_t = \sigma_2(\boldsymbol{o}_t \otimes \boldsymbol{c}_t) \tag{17}$$

where $\boldsymbol{i}_t$, $\boldsymbol{o}_t$,$\boldsymbol{f}_t$ are respectively the input gate vectors, the output gate vectors, and the forget gate vectors. $\boldsymbol{W}$, $\boldsymbol{U}$, $\boldsymbol{b}$ are respectively weight matrixes and bias vectors for each gate. $\otimes$ represents the element-wise multiplication. We set the activation function σ_1 as the sigmoid function. σ_2 is the hyperbolic tangent function (Tanh). Both the forward and backward recurrent operations are in a similar way of computation according to the equations above.

3.4 Style embedding module

The emotional expression of music is highly related to the style. Pop music focuses on the expression of the human voice, while rock music focuses on the expression of the instrument. The dataset we used covers 8 musical styles, specifically classical, rock, blues, folk, jazz, country, electronic, and pop. To take advantage of the style information in metadata, we design a style embedding module. Given the flatten output from BLSTM module is $\boldsymbol{V}_{FH}$, and this

vector is converted to $\boldsymbol{V}_{\boldsymbol{FH}}^{'}$ by a fully connected layer. The style vector $\boldsymbol{V}_{\boldsymbol{s}}$ is generated by one-hot encoding. To control the range of the vector, $\boldsymbol{V}_{\boldsymbol{FH}}^{'}$ is normalized by a min-max scaler, and the calculation is as:

$$\boldsymbol{V}_{\boldsymbol{FH}}^{'} = \frac{\boldsymbol{V}_{\boldsymbol{FH}}^{'} - \min\left(\boldsymbol{V}_{\boldsymbol{FH}}^{'}\right)}{\max\left(\boldsymbol{V}_{\boldsymbol{FH}}^{'}\right) - \min\left(\boldsymbol{V}_{\boldsymbol{FH}}^{'}\right)} \tag{18}$$

where min, max are the operations to calculate the minimum and maximum value of $\boldsymbol{V}_{\boldsymbol{FH}}^{'}$. Then the concatenated vector $\boldsymbol{V}_{\boldsymbol{E}}$ can be calculated as:

$$\boldsymbol{V}_{\boldsymbol{E}} = \boldsymbol{V}_{\boldsymbol{FH}}^{'} \oplus \boldsymbol{V}_{\boldsymbol{S}} \tag{19}$$

where $\oplus$ is the concatenation operation. Finally, the processed vector will be input into the later fully connected layer to construct the regression module. The style embedding module is shown in Fig. 6.

The average deviation like Mean Absolute Error (MAE), Mean Squared Error (MSE) [54] or ϵ-insensitive MAE [18] and the correlation coefficient like Pearson's Correlation Coefficient (PCC) [39, 45] are often used as the loss function in previous continuous emotion recognition tasks. The performance of the networks is determined by the minimization of average deviation criteria but the maximization of correlation coefficients. The task "Emotion in Music" in MediaEval has been a benchmark in MER since 2013 [1], in which CCC is one of the metrics for continuous emotion recognition. Another benchmark in MER, which is the Multimodal Affect Recognition Sub-Challenge (MASC) of Audio-Visual Emotion Challenge (AVEC)[4] since 2015, also utilize CCC as the criterion. Therefore, CCC is widely applied in continuous emotion recognition because the characteristics of both the two types of metrics have been considered. It has been proven to perform better than other metrics in previous researches.

Thus, we also utilize CCC loss as the objective function in our experiments. $\boldsymbol{Y_b} = \left[y_b^1, y_b^2, \ldots\ldots, y_b^{N_b}\right]$ is the vector of labels of the mini-batch b and N_b is the number of music pieces within mini-batch b. $\widehat{\boldsymbol{Y}}_{\boldsymbol{b}} = \left[\widehat{y}_b^1, \widehat{y}_b^2, \ldots\ldots, \widehat{y}_b^{N_b}\right]$ is the prediction vector. The PCC of each mini-batch of music piece is computed as:

$$PCC_b = \frac{\sum_{i=1}^{N_b}\left(y_b^i - \overline{Y_b}\right)\left(\widehat{y}_b^{\,i} - \overline{\widehat{Y}_b}\right)}{\sqrt{\sum_{i=1}^{N_b}\left(y_b^i - \overline{Y_b}\right)^2 \sum_{i=1}^{N_b}\left(\widehat{y}_b^{\,i} - \overline{\widehat{Y}_b}\right)^2}} \tag{20}$$

where PCC_b denotes the PCC of the music piece mini-batch b. $y_b^i, \widehat{y}_b^{\,i}$ are respectively the ground-truth and the prediction of music piece i. $\overline{Y}_b, \overline{\widehat{Y}}_b$ are separately the mean value of the labels and the predictions. The CCC loss mini-batch b is calculated as:

$$CCC_b = \frac{2PCC_b\sigma_{Y_b}\sigma_{\widehat{Y}_b}}{{\sigma_{Y_b}}^2 + {\sigma_{\widehat{Y}_b}}^2 + \left(\mu_{\widehat{Y}_b} - \mu_{Y_b}\right)^2} \tag{21}$$

[4] AVEC. Available: https://avec-db.sspnet.eu/

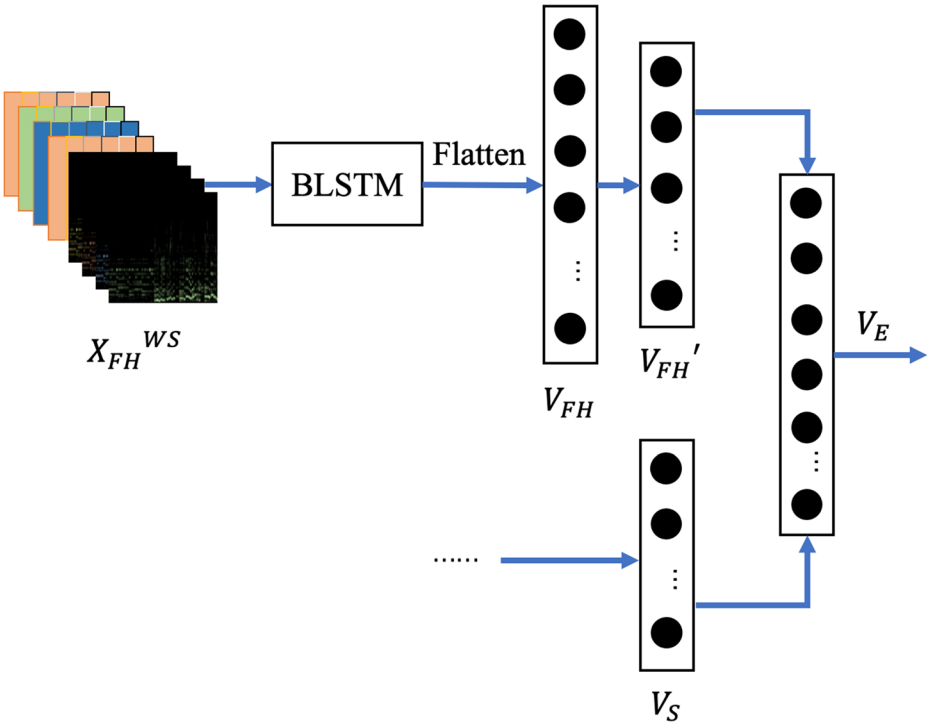

Fig. 6 Style embedding module. The process of style embedding

$$Loss_b = 1 - CCC_b \tag{22}$$

where $\mu_{Y_b} = \overline{Y_b}, \mu_{\widehat{Y_b}} = \overline{\widehat{Y}}_b$ are respectively the mean value of labels and predictions.$\sigma_{Y_b}{}^2 = \sum_{i=1}^{N_b}\left(y_b^i - \mu_{Y_b}\right)^2/(\mathrm{N}_b - 1)$, $\sigma_{\widehat{Y_b}}{}^2 = \sum_{i=1}^{N_b}\left(\widehat{y}_b^i - \mu_{\widehat{Y_b}}\right)^2/(N_b - 1)$ are separately the variance of labels and predictions. Therefore, the average batch loss is calculated as the average CCC loss:

$$Loss_{Total} = \frac{\sum_{i=1}^{N_{batch}} Loss_i}{N_{batch}} \tag{23}$$

where $Loss_{Total}$ denotes the average training loss of all the music pieces. $Loss_i$ is the CCC loss of mini-batch N_b is the total size of the total mini-batches.

4 Experiments and results

4.1 Datasets

We use two public continuous music emotion datasets to train and validate the proposed model. To guarantee the input conveyed to the networks in the same format, the original audio is transformed to the same down sampled rate 22,050 Hz, WAV format, mono channel, and 16 bits PCM encoding. The details of the two datasets are as follows.

DEAM[5]: 1802 songs with valence and arousal annotations. The dataset is established by Aljanaki et al. [1] and is the largest benchmark in continuous music emotion recognition. However, the consistency of annotations is low due to the different periods of annotating and subjectivity. Thus, we choose the subset of DEAM called 1000 songs[6] (744 segments of 45 s) with higher consistency for the same period annotation according to Soleymani et al. [37]. In our experiments, we split the dataset into three parts, that is, 434 pieces for training, 155 pieces for validating, and 155 pieces for testing to verify the performance of MCAN. To match the per second (from 15,000 ms to 44,500 ms) dynamic valence and arousal annotations scaled in [−1, 1], we split each original raw audio into 60 subsegments (500 ms per) without overlapping discarding 15,000 ms from the beginning of every music segment.

PMEmo[7]: 794 popular songs with valence and arousal annotations. This dataset is established by Zhang et al. [53], in which lyrics and comments are added in 2019 to make this dataset much more suitable for cross-modal research. In our experiments, we utilize this dataset in mono-modality to validate the generalizability of MCAN. The value of dynamic annotations in PMEmo is in the range [0, 1]. We remove the samples without sufficient annotations and the total chosen size is 206 pieces (164 for training, 21 for validating, and 21for testing). Additionally, we filter out all the chosen samples with the dynamic annotations in the range of 15,000 ms to 45,000 ms. Consequently, we split the chosen raw audio into 60 subsegments (500 ms per) without overlapping to match the dynamic annotations.

4.2 Experimental settings

To prompt the performance of MCAN, we apply the mixup [52] technique before the spectrograms being transferred to the first convolutional layer, in which we set α as 0.2 to limit the weight parameter ω mentioned in Section 3. The batch size is set as 15, not only because of the limitation of the experimental environment, but also to facilitate the calculation of $Loss_{per}$. We use Adam optimizer to modify the model while training. The learning rate is 0.001 in the beginning, which will decrease by 1/10 every 10 epochs. The dropout technique is used after BLSTM and context-attention layers to prevent overfitting on the training dataset, where the ratio of dropout is respectively 0.3 and 0.2. L2 regularization is also applied to the weights of the fully connected layer. Our experiments are performed in the PyTorch framework written in python and are carried out on NVIDIA GeForce GTX 1080 with 32 GB onboard memory.

5 Results and evaluation

In this section, we introduce the experiments conducted in MCAN. Metrics we choose to evaluate the performance of MCAN are the Root Mean Squared Error (RMSE), MAE, and CCC. According to the distribution of annotations, they obey the normal distribution, which is the premise to use PCC. Thus, CCC that is based on PCC can be used as the evaluation indicator. RMSE enables the network to converge faster while MAE is more sensitive to the outliers. Thus, we use both RMSE and MAE as auxiliary metrics in our experiments. CCC is

[5] DEAM. Available: http://cvml.unige.ch/databases/DEAM/
[6] 1000 Songs. Available: http://cvml.unige.ch/databases/emoMusic/
[7] PMEmo. Available: http://www.next.zju.edu.cn/research/pmemo/amp/

utilized for the main analysis of the performance of MCAN while the hypothesis is accepted within the confidence ($p < 0.05$) of the T-test. RMSE and MAE are employed as judgments while the CCC fails to satisfy the T-test.

5.1 Parameter adjustment

In order to observe the influence of different parameters on the experimental results and obtain better experimental results, we conduct parameter adjustment experiments on DEAM dataset. The parameters observed in these experiments are α and β, where α is the parameter introduced in section 3.1 to control the proportion of reconstructed samples and β is the weight parameter introduced in section 3.3 to control the summation of features. The results on testing dataset of the adjustment experiments are shown in Figs. 7 and 8.

In the adjustment experiment of parameter α, other variables are kept unchanged. As is shown in Fig. 7, the optimal solution can be obtained on all metrics when α is set between 0.1 and 0.2, which means that introducing this parameter can effectively control the performance of the model. If the sample reconstruction technique is not applied, that is, when $\alpha = 0$, the results will be worse. It illustrates the importance of sample reconstruction. Furthermore, when α is over 0.2, the results get worse as parameter α increases. It shows that too much sample reconstruction will disturb the learning ability of the model, so it is necessary to control the proportion of reconstructed samples. This parameter is set as 0.18 in our experiment. In general, the introduction of parameter α can help to control the number of reconstructed samples, and has a great contribution to the improvement of the results.

In the adjustment experiment of parameter β, other variables are kept unchanged. As is shown in Fig. 8, the optimal solution can be obtained on all metrics when β is set between 0.4 and 0.6, which indicates the validity of the weighted attention module. If the attention mechanism is not applied, that is, when $\beta = 0$, the results will be worse. It verifies the effectiveness of introducing self-attention mechanism for fine-grained processing of features. Furthermore, when β exceeds 0.6, the results get worse as parameter β increases. This illustrates that it is necessary to control the weight while performing feature fusion on original

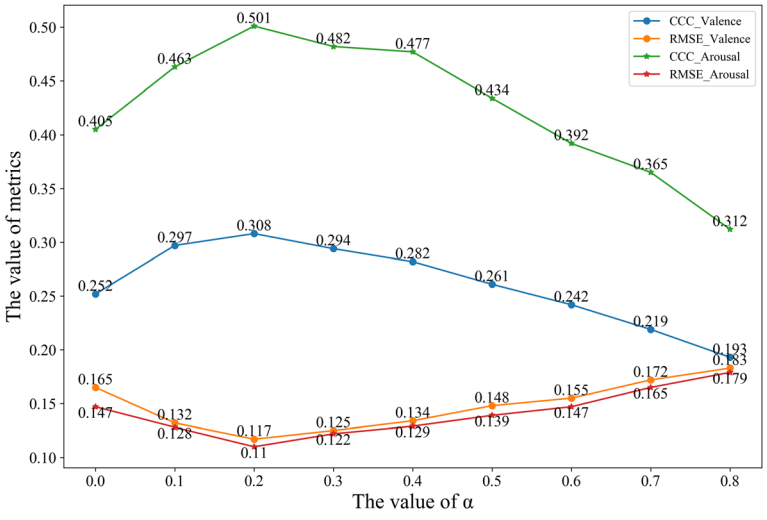

Fig. 7 The value of metric varies by parameter α in *valence* and *arousal* dimensions

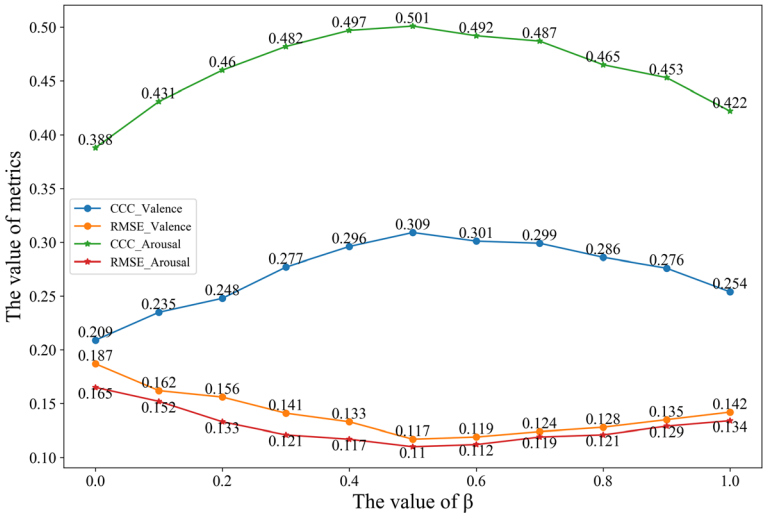

Fig. 8 The value of metric varies by parameter β in *valence* and *arousal* dimensions

features and processed ones. This parameter is set as 0.49 in our experiment. To sum up, the introduction of parameter β can help to balance the influence of different features, and can effectively control this situation to improve the performance of the model.

5.2 Ablation study

To validate whether the techniques utilized in our model can improve the performance, we selectively remove some operations and conduct a series of comparative experiments on DEAM dataset, which mainly contains the following four cases:

Case1: The sample reconstruction technique is removed.
Case2: The branch of handcrafted features is removed, which means no feature augmentation. In this case, the sample reconstruction is only aimed at the filter bank output.
Case3: The weighted attention module is removed, which means no self-attention mechanism and no BLSTM module.
Case4: The style embedding module is removed.

The above-ablated models are named *Case1*, *Case2*, *Case3*, and *Case4* respectively. Tables 1 and 2 show the results of the testing dataset on DEAM of different versions. MCAN performs better than any of the ablation versions. Modification tricks that we apply in our model are helpful for the prediction. Take the results of CCC as an example, the values of CCC ($p < 0.05$) in valence are 0.155 in *Case2* and 0.185 in *Case3*. Comparing the above values to the result 0.309 of MCAN, the value of CCC has improved significantly, which suggests that the feature augmentation module and the weighted attention module are both effective in this task. The adjustment experiment of parameter β has also confirmed the contribution of feature fusion in these modules. The construction of these modules can contribute to extract features highly related to the expression of emotions. The values of CCC ($p < 0.05$) in arousal are evenly 0.379 in *Case1* and 0.402 in *Case4*. Comparing the above values to the result 0.502 of

Table 1 The results of the ablation study in *valence* on testing datasets of DEAM

	CCC	RMSE	MAE
MCAN	**0.309±0.358**	**0.112±0.010**	**0.811±0.009**
Case1	0.252±0.402	0.165±0.066	0.992±0.040
Case2	0.155±0.331	0.223±0.125	1.089±0.117
Case3	0.185±0.244	0.201±0.097	1.095±0.102
Case4	0.210±0.421	0.155±0.132	0.983±0.127

The bold MCAN means that MCAN is the complete version of our method. The bold values of different metrics mean that MCAN performs better than other ablation ones in valence dimension. The higher the value of the CCC indicator, the better the performance. Lower values for the RMSE and MAE metrics indicate better performance

MCAN, the value of CCC has increased, which illustrates that the sample reconstruction technique and the style embedding module can also help to prompt the precision of predictions and they have similar contributions according to the close values. The former makes the network more robust by adding noisy samples, while the latter makes the network more sensitive in processing of details by embedding style vectors. Furthermore, the adjustment experiment of parameter α has also shows the contribution of sample reconstruction technique. Comparing different ablation versions, it can be seen that the feature augmentation and the weighted attention module with higher values of CCC are the main techniques that help improving the performance of MCAN. The results of RMSE and MAE have an analogous tendency and have decreased a lot while applying MCAN, which further demonstrates the usefulness of the adopted algorithms.

5.3 Results on DEAM and PMEmo

Figure 9 illustrates the variation of training CCC loss on DEAM and PMEmo datasets in valence and arousal dimensions. The convergence occurs in epoch among the range of (60, 80) on both two datasets and also in different dimensions of emtion, which demonstrates that the perception of emotion is of great commonality and MCAN is able to perceive this commonality. The training loss has close downward trend in two different datasets, which suggests that MCAN is effective in the task of numerical prediction for emotion in music. MCAN has similar performance on the same type of dataset, indicating that it has strong ability of generalization.

Tables 3 and 4 represent the values of metrics while testing, which demonstrates the relationship between different perceived emotions of humans and validates the generalizability of MCAN. For example, the values of CCC ($p < 0.05$) reach evenly 0.309 in valence and 0.502 in arousal on the DEAM testing dataset. The values are 0.215 and 0.401 on the PMEmo

Table 2 The results of the ablation study in *arousal* on testing datasets of DEAM

	CCC	**RMSE**	**MAE**
MCAN	**0.502±0.287**	**0.109±0.011**	**0.764±0.023**
Case1	0.405±0.314	0.147±0.072	0.842±0.082
Case2	0.337±0.329	0.210±0.098	0.997±0.073
Case3	0.379±0.298	0.192±0.025	1.008±0.017
Case4	0.402±0.324	0.138±0.102	0.934±0.124

The bold MCAN means that MCAN is the complete version of our method. The bold values of different metrics mean that MCAN performs better than other ablation ones in arousal dimension. The higher the value of the CCC indicator, the better the performance. Lower values for the RMSE and MAE metrics indicate better performance

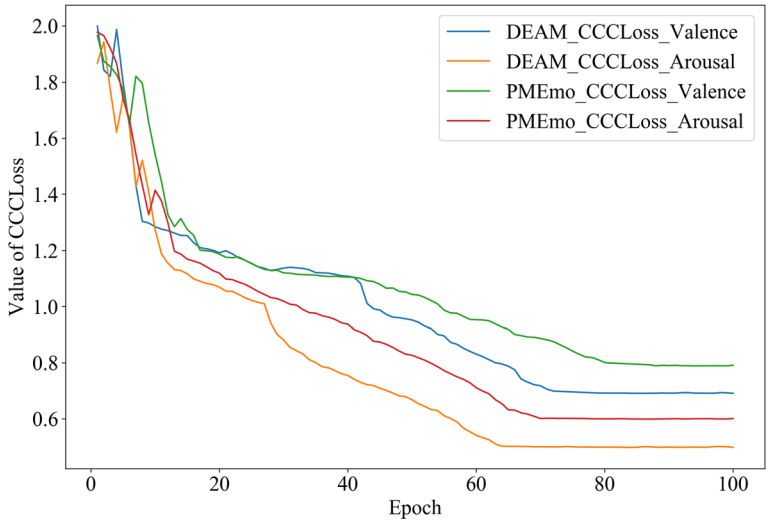

Fig. 9 Running loss $Loss_{total}$ varies by epoch in *valence* and *arousal* dimensions on DEAM and PMEmo training datasets

testing dataset. From these values, it can be seen that the results on DEAM and PMEmo are in the similar range and tendency in both valence and arousal dimensions. The two datasets are annotated by different groups of people, under the condition of which the results mentioned above demonstrate that the ability of humans to perceive emotion is close. The results also suggests that MCAN is well applicable for continuous MER on similar datasets with dynamic annotations. In other words, MCAN has the ability of generalization and is very stable in different situations.

5.4 Comparison with other methods

We compare MCAN with five methods, and the details of these methods are as follows:

> *The top-three baselines:* BLSTM-RNN, BLSTM-ELM, and Deep LSTM-RNN [1], which are applied in the "Emotion in Music" task at MediaEval. All of these have received desirable results in continuous MER, and these methods are regarded as the benchmark ones in MER tasks.

Table 3 The representation of the results in *valence* on different testing datasets

	CCC	RMSE	MAE
DEAM	0.309±0.358	0.112±0.010	0.811±0.009
PMEmo	0.215±0.265	0.144±0.102	0.892±0.030

Table 4 The representation of the results in *arousal* on different testing datasets

	CCC	RMSE	MAE
DEAM	0.502±0.287	0.109±0.011	0.764±0.023
PMEmo	0.401±0.315	0.135±0.057	0.901±0.103

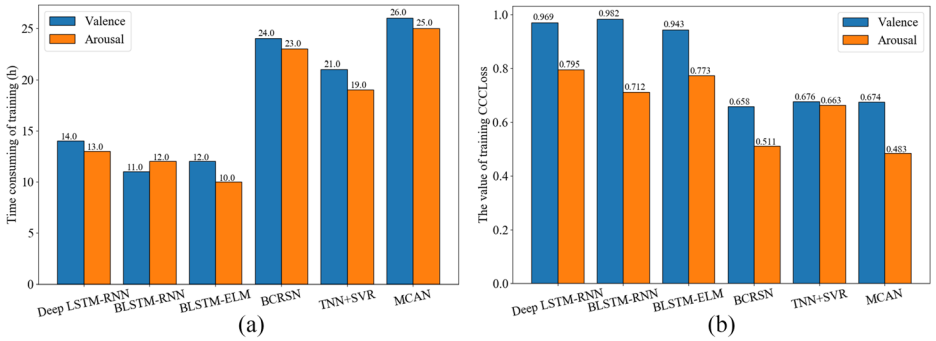

Fig. 10 Comparison of performance on DEAM training dataset. **a**: Time-consuming training per epoch comparison chart of different methods. **b**: $Loss_{total}$ comparison chart of different methods

The two state-of-the-art methods: The first one is the method called BCRSN [12], in which CNNs and LSTM module are applied together to construct the method. The second one is a novel method based on the combination of an effective dimensionality reduction algorithm and SVR [8], which can be descripted as TNN + SVR.

Tables 5 and 6 represent the performance of different models on the DEAM and PMEmo testing dataset. Compared to the top-three baselines, MCAN outperforms these models for the ability to extract salient features through original spectrograms. The values of CCC have improved significantly and the values of RMSE and MAE have decreased obviously while applying MCAN, which indicates that the proposed MCAN can effectively extract emotion-related features and contribute to better performance. Taking the results in the ablation experiments into consideration, the combined effect of the sample reconstruction technique, the feature augmentation, the weighted attention module and the style embedding module makes MCAN perform well. From the perspective of data representation, there are sufficient data representation in details in our method. The handling of details makes MCAN more stable and accurate.

Table 5 The comparison of the performance in *valence* on different models and testing datasets

		CCC	RMSE	MAE
DEAM	MCAN **(ours)**	0.309±0.358	**0.112±0.010**	**0.811±0.009**
	Deep LSTM-RNN	0.012±0.421	0.256±0.201	1.040±0.102
	BLSTM-RNN	0.001±0.240	0.314±0.098	1.019±0.097
	BLSTM-ELM	0.049±0.425	0.320±0.186	1.108±0.152
	BCRSN	0.295±0.330	0.123±0.103	0.925±0.097
	TNN+SVR	**0.311±0.212**	0.125±0.074	0.954±0.079
PMEmo	MCAN **(ours)**	**0.215±0.265**	**0.144±0.102**	0.892±0.030
	Deep LSTM-RNN	0.009±0.105	0.310±0.042	1.127±0.098
	BLSTM-RNN	0.001±0.536	0.305±0.099	1.201±0.102
	BLSTM-ELM	0.032±0.378	0.241±0.101	1.039±0.099
	BCRSN	0.209±0.415	0.158±0.090	0.959±0.102
	TNN+SVR	0.213±0.255	0.149±0.101	**0.889±0.059**

The bold entries mean better performance in valence dimension. The higher the value of the CCC indicator, the better the performance. Lower values for the RMSE and MAE metrics indicate better performance

Table 6 The comparison of the performance in *arousal* on different models and testing datasets

		CCC	RMSE	MAE
DEAM	MCAN **(ours)**	**0.502±0.287**	0.109±0.011	**0.764±0.023**
	Deep LSTM-RNN	0.195±0.347	0.215±0.057	1.002±0.007
	BLSTM-RNN	0.210±0.365	0.231±0.104	1.108±0.102
	BLSTM-ELM	0.213±0.423	0.179±0.103	1.078±0.097
	BCRSN	0.472±0.302	**0.101±0.100**	0.875±0.022
	TNN+SVR	0.279±0.412	0.185±0.081	0.914±0.103
PMEmo	MCAN **(ours)**	**0.401±0.315**	0.135±0.057	**0.901±0.103**
	Deep LSTM-RNN	0.165±0.298	0.298±0.050	1.295±0.040
	BLSTM-RNN	0.197±0.320	0.218±0.097	1.215±0.045
	BLSTM-ELM	0.202±0.254	0.185±0.105	1.183±0.100
	BCRSN	0.320±0.329	**0.132±0.059**	0.997±0.057
	TNN+SVR	0.315±0.229	0.145±0.022	1.002±0.104

The bold entries mean better performance in arousal dimension. The higher the value of the CCC indicator, the better the performance. Lower values for the RMSE and MAE metrics indicate better performance

Furthermore, we also compare MCAN with the state-of-the-art methods. As is shown in Tables 5 and 6, MCAN is not the best in all conditions. As for the DEAM dataset, TNN + SVR performs better than MCAN on the evaluation of CCC in valence dimension, and BCRSN performs better than MCAN on the evaluation of RMSE in arousal dimension. As for PMEmo dataset, TNN + SVR performs better than MCAN on the evaluation of MAE in valence dimension, and BCRSN performs better than MCAN on the evaluation of RMSE in arousal dimension. However, compared with the values of TNN + SVR or BCRSN on these indicators, the values of MCAN's are not far behind those ones, which are less than 1%. From another point of view, compared with BCRSN, the values of other indicators of MCAN have been improved about 8%, and compared with TNN + SVR, the values of other indicators of MCAN have also been improved about 9%. Compared with BCRSN and TNN + SVR, the data representation of MCAN is more adequate and the process of feature extraction is finer. The automatic feature extraction through spectrogram and the handcrafted feature extraction are combined together in MCAN, which makes it perform better than the two state-of-the-art methods. Furthermore, the sample reconstruction technique can help to strengthen the stability of the network. The feature augmentation and the weighted attention module can help to extract salient features. The style embedding module can help to enhance the learning capability of the network. Overall, MCAN performs better than all methods compared.

Figure 10(a) shows the training time of different methods evenly on the DEAM training dataset and Fig. 10(b) represents the average CCC loss on the DEAM training dataset. It can be seen that MCAN outperforms other methods in both valence and arousal dimensions but it costs a relatively long time-consuming. Compared with the top-three baselines, the performance of MACN is improved significantly, which indicates that the salient features extracted from spectrograms and handcrafted features are important in MER. Compared with the two state-of-the-art methods, the CCC loss drop obviously, which demonstrates that the detailed feature extraction is essential in MER. Comparing the top-three baselines and other methods, it is obvious that the architectures combined with CNNs and LSTM can extract emotion-related semantic features better. These features outperform those extracted by basic networks. Additionally, the results of MACN illustrate that the combination of sample reconstruction, feature augmentation, weighted attention module and style embedding module can achieve state-of-the-art performance in continuous MER.

6 Conclusion and future work

In this paper, we propose a novel MCAN to extract salient emotion-related feature maps from filter bank output and handcrafted features. The sample reconstruction technique can strengthen the stability of the network and contribute to better performance. The feature augmentation module makes the process of feature extraction more fine-grained, and it is helpful for extracting salient features. The weighted attention module is useful for controlling the focus of the network and processing the sequential information in details. Especially, the style embedding can help to enhance the learning ability of the network. The superior results on the benchmark dataset DEAM and the new proposed PMEmo represent the effectiveness and versatility of MCAN.

However, there are shortcomings of the proposed MCAN. Firstly, the training time of our model is relatively long. Secondly, our model is unstable while dealing with drastic short-term variance in emotions. Finally, our model has flaws in handling outliers. In the future, we will concentrate on simplifying the structures of our method and reducing the parameters contained in models. Additionally, we will consider fusing information of extra modality (e.g., electro-encephalogram (EEG), lyric [24, 47]) to achieve more accurate prediction and improve the stability of the method.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11042-022-13577-6.

Code availability Custom code is not available without restriction.

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Meixian Zhang and Yonghua Zhu. The first draft of the manuscript was written by Meixian Zhang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability Data and material are fully available without restriction.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Aljanaki A, Yang YH, Soleymani M (2017) Developing a benchmark for emotional analysis of music. PLoS One 12(3):e0173392
2. Amiriparian S, Gerczuk M et al (2019) Emotion and themes recognition in music utilizing convolutional and recurrent neural networks. In: Proceedings of the MediaEval 2019 workshop

3. Bharti D, Kukana P (2020) A hybrid machine learning model for emotion recognition from speech signals. In: Proceedings of the International Conference on Smart Electronics and Communication, pp. 491–496
4. Bogdanov D, Wack N et al (2013) ESSENTIA: an audio analysis library for music information retrieval. In: Proceedings of the 14th International Society for Music Information Retrieval Conference, pp 493–498
5. Bogdanov D, Porter A, Tovstogan P et al (2019) MediaEval 2019: emotion and theme recognition in music using Jamendo. In: Proceedings of the MediaEval 2019 workshop
6. Cabrera D et al (1999) Psysound: a computer program for psychoacoustical analysis. In: Proceedings of the Australian acoustical society conference, 24: 47–54
7. Chen S, Jin Q (2015) Multi-modal dimensional emotion recognition using recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp 49–56
8. Cheuk KW, Luo YJ et al Regression-based music emotion prediction using triplet neural networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks, pp 1–7
9. Deshpande H, Singh R, Nam U (2001) Classification of music signals in the visual domain. Proceedings of the COST G-6 Conference on Digital Audio Effects, 3(1): 1–4
10. Dieleman S, Brakel P, Schrauwen B (2011) Audio-based music classification with a pretrained convolutional network. In: Proceedings of the 12th International Symposium on Music Information Retrieval, pp 669–674
11. Dingle GA, Kelly PJ et al (2015) The influence of music on emotions and cravings in clients in addiction treatment: a study of two clinical samples. The Arts in Psychotherapy 45:18–25
12. Dong Y, Yang X, Zhao X, Li J (2019) Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. IEEE Transactions on Multimedia 21(12):3150–3163
13. Florence SM, Uma M (2020) Emotional detection and music recommendation system based on user facial expression. IOP Conference Series: Materials Science and Engineering 912(6):062007
14. Grekow J (2018) From content-based music emotion recognition to rmotion maps of musical pieces. Polish Academy of Science, Polish Warsaw
15. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
16. Huan RH, Shu J et al (2020) Video multimodal emotion recognition based on Bi-GRU and attention fusion. Multimedia Tools and Applications, pp 1–28
17. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708
18. Huang J, Li Y, Tao J et al (2018) Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In: Proceedings of the 8th international workshop on audio/visual emotion challenge, pp 57–64
19. Hung HT, Chen YH et al (2019) MediaEval 2019 emotion and theme recognition task: a VQ-VAE based approach. In: Proceedings of the MediaEval 2019 Workshop
20. Huo Y, Yao H et al (2020) Soul dancer: emotion-based human action generation. ACM Transactions on Multimedia Computing Communication and Application 15(3s):1–19
21. S. Jun, H. Hwang (2009) A fuzzy inference-based music emotion recognition system. In: Proceedings of the 5th International Conference on Visual Information Engineering, pp 673–677
22. Lartillot O, Toiviainen P (2007) MIR in matlab (II): a toolbox for musical feature extraction from audio. In: Proceedings of the International Conference on Music Information Retrieval
23. Li T, Ogihara M (2003) Detecting emotion in music. In: International Symposium on Music Information Retrieval, pp. 239–240
24. Li Y, Zheng W (2021) Emotion recognition and regulation based on stacked sparse auto-encoder network and personalized reconfigurable music. Mathematics 9(6):593
25. Li X et al (2016) A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 544–548
26. MacDorman OC, Stuart H, Karl F (2007) Automatic emotion prediction of song excerpts: index construction, algorithm design, and empirical comparison. Journal of New Music Research 36(4):281–299
27. Malik M, Adavanne S, Drossos K et al (2017) Stacked convolutional and recurrent neural networks for music emotion recognition. In: Proceedings of the 14th sound music computing conference, pp 208–213
28. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia 16(8):2203–2213
29. Mayerl M, Vötter M et al (2019) Recognizing song mood and theme using convolutional recurrent neural networks. In: Proceedings of the MediaEval 2019 workshop
30. Orjesek R, Jarina R, Chmulik M et al (2019) DNN based music emotion recognition from raw audio signal. In: proceeding of the 29th international conference Radioelektronika, pp 1-4

31. Patra BG, Das D, Bandyopadhyay S (2013) Unsupervised approach to Hindi music mood classification. In: Proceedings of the Mining Intelligence and Knowledge Exploration, pp. 62–69
32. Pons J, Nieto O, Prockup M et al (2018) End-to-end learning for music audio tagging at scale. In: Proceedings of the 12th international symposium/conference on music information retrieval, pp 637–644
33. Sangnark S, Lertwatechakul M, Benjangkaprasert C (2018) Thai music emotion recognition based on western music. J Phys Conf Ser 1195(1):012009
34. Sarkar R, Choudhury S, Dutta S, Roy A, Saha SK (2020) Recognition of emotion in music based on deep convolutional neural network. Multimed Tools Appl 79:765–783
35. Schmidt EM, Kim YE (2010) Prediction of time-varying musical mood distributions from audio. In: Proceedings of the International Society of Music Information Retrieval Conference, pp. 465–470
36. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations
37. Soleymani MM, Caro MN (2013) 1000 Songs for emotional analysis of music. In: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia, pp 1–6
38. Sukhavasi M, Adapa S (2019) Music theme recognition using CNN and self-attention. In: Proceedings of the MediaEval 2019 Workshop
39. Sun L, Lian Z et al (2020) Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In: Proceedings of the 1st International on Multimedia Sentiment Analysis in Real-life Media Challenge and Workshop, pp 27–34
40. Thayer RE (1989) The biopsychology of mood and arousal. Personal Individ Differ 11(9):993–993
41. Tzanetakis G, Cook P (2000) Marsyas: a framework for audio analysis. Organized Sound 4(3):169–177
42. Wang Y, Sun S (2019) Emotion recognition for internet music by multiple classifiers. In: Proceedings of the IEEE/ACIS 18th International Conference on Computer and Information Science, pp 262–265
43. Wang JC, Yang YH et al (2012) The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In: Proceedings of the 20th ACM international conference on multimedia. pp 89–98
44. Wang Y, Wu J et al (2019) Multi-attention fusion network for video-based emotion recognition. In: Proceedings of the International Conference on Multimodal Interaction, pp. 595–601
45. Weninger F, Ringeval F, Marchi E et al (2016) Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In: Proceedings of the International Joint Conference Artificial Intelligence, pp. 2196–2202
46. Wu B, Zhong E, Horner A et al (2014) Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In: Proceedings of the 22nd ACM international conference on multimedia, pp 117–126
47. Yang J (2021) A novel music emotion recognition model using neural network technology. Front Psychol 12:760060
48. Yang YH, Chen HH (2011) Music emotion recognition. CRC Press, Boca Raton
49. Yang YH, Liu CC, Chen HH (2006) Music emotion classification: a fuzzy approach. In: Proceedings of the 14th ACM International Conference on Multimedia, pp 81–84
50. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 16(2):448–457
51. Yang X, Dong Y, Li J (2018) Review of data features-based music emotion recognition methods. Multimedia Systems 24(4):365–389
52. Zhang H, Cisse M et al (2018) Mixup: beyond empirical risk minimization. In: Proceedings of the 6th international conference on learning representations
53. Zhang K, Zhang H et al (2018) The PMEmo dataset for music emotion recognition. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval, pp 135-142
54. Zhao J, Li R et al (2019) Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. In: Proceedings of the 9th international workshop on audio/visual emotion challenge, pp 37–45
55. Zhao S, Li Y et al (2020) Emotion-based end-to-end matching between image and music in valence-arousal space. In: Proceedings of the 28th ACM international conference on multimedia, pp 2945-2954