

Extracting Emotional Feature Changes in College Music Performances via Multi-Channel Convolutional Attention Mechanism

Qin Kang¹

¹College of Music, Hefei Normal University, Hefei 230601, Anhui, China

Corresponding author: Qin Kang (e-mail: Kangqin198005@163.com)

ABSTRACT To address the limitations of traditional emotional-feature extraction methods for musical performances, including low accuracy, inadequate temporal-continuity modeling, and weak differentiation of subtle emotional changes, this study proposes a multi-channel convolutional attention-based method for extracting emotional feature variations. A multimodal data acquisition system for university music performances is first constructed to collect audio data, including pitch, rhythm, and dynamics, and visual data, including facial expressions and body movements, from vocal and instrumental performances. A multi-channel convolutional neural network is then designed for multidimensional feature extraction from heterogeneous data. The audio channel uses 1D-CNN to capture time-domain and frequencydomain features, while the visual channel combines 2D-CNN and 3D-CNN to extract spatial and temporal sequence features. A hybrid attention mechanism integrating channel attention and temporal attention is introduced to adaptively weight key emotional features and enhance the capture of dynamic emotional changes. Experiments on the self-built UMPED dataset and the public MusicNet dataset evaluate emotion-feature similarity, classification accuracy, and temporal emotion-change fitting, verifying the effectiveness of the proposed multimodal signal-processing framework.

INDEX TERMS multi-channel convolution, attention mechanism, music performance, emotion feature extraction, signal processing

I. INTRODUCTION

A. RESEARCH BACKGROUND

UNIVERSITY music performances serve as a vital practical component of music education, comprehensively reflecting students' artistic literacy, performance skills, and emotional expression capabilities [1]. As the core essence of these performances, the accuracy and impact of musical emotion conveyance directly determine performance quality [2]. Against the backdrop of digital transformation in music education, objectively extracting emotional feature variations during performances through technological means not only provides teachers with quantifiable teaching evaluation criteria and helps students pinpoint weaknesses in emotional expression, but also offers data support for assessing performance quality and cultivating artistic talent [3], [4]. The integration of multi-channel convolutional and attention mechanisms offers a novel approach to addressing these challenges. Multi-channel structures enable parallel feature extraction from heterogeneous data, while attention mechanisms adaptively focus on critical emotional information [5], [6]. Therefore, researching emotional feature extraction

during university music performances using multi-channel convolutional attention mechanisms holds significant theoretical and practical value for advancing intelligent music education and enhancing the objectivity and precision of performance quality assessment.

B. CURRENT RESEARCH STATUS

1) Research on Music Emotional Feature Extraction Technology

International research began earlier, forming preliminary multimodal fusion techniques. Bello et al. [7] constructed an emotion classification model based on Mel-frequency cepstral coefficients (MFCC) and rhythmic features from audio signals, achieving recognition of basic emotions like joy and sadness. Juslin et al. [8] integrated facial expressions and body movement data to propose a visual emotion feature extraction framework, enhancing the comprehensiveness of emotion recognition. However, most international studies focus on general music emotion recognition and have not optimized models for the specialized nature of university performances or the subtlety of emotional expression [9]–

[11].

2) Application of Attention Mechanisms in Affective Computing

Attention mechanisms have been widely adopted in affective computing due to their ability to amplify key information while suppressing redundant noise. Vaswani et al. [12] introduced the self-attention mechanism, providing a foundational framework for affective feature extraction. Wang et al. [13] designed a channel attention module that adaptively adjusts weights across audio frequency channels, enhancing the discriminative power of affective features. Li et al. [14] introduced a temporal attention mechanism to focus on critical time intervals with intense emotional fluctuations. However, existing attention models lack targeted optimization for multimodal, long-sequence, and fine-grained emotional features in university music performance scenarios, making it difficult to meet the precise demands of performance emotion extraction.

C. RESEARCH CONTENT

1. Multimodal Data Collection and Dataset Construction for University Music Performances: Collect audio and visual data from vocal (bel canto, ethnic, pop), instrumental (piano, violin, guzheng) performances, annotate emotional types and transition points, and construct the UMPED dataset;
2. Design of Multi-Channel Convolutional Feature Extraction Network: Develop dedicated convolutional channels for audio and visual modalities to enable parallel extraction of multi-dimensional features;
3. Hybrid Attention Mechanism Construction: Integrate channel attention and temporal attention with adaptive weighting strategies to enhance capture of key emotional features and dynamic changes;
4. Experimental Validation and Performance Evaluation: Validate method effectiveness using self-built and public datasets, assessing dimensions such as emotion classification accuracy and temporal fit.

II. THEORETICAL FOUNDATIONS

A. EMOTIONAL FEATURE ANALYSIS IN UNIVERSITY MUSIC PERFORMANCES

1) Audio Emotional Features

Audio emotional features serve as the core vehicle for emotional expression, encompassing acoustic and musicological characteristics as detailed in Table 1. Single audio features cannot independently determine emotions. The model achieves emotion differentiation through multi-feature fusion: for example, “high STE + high F0 + major mode” corresponds to excitement, while “high STE + low F0 + minor mode” corresponds to anger. This avoids misjudgment caused by relying on individual features.

2) Visual Emotional Features

Visual emotional features are conveyed through an actor’s external expressions, encompassing facial expressions and body movement characteristics:

- Facial expression features: Extracted based on facial landmarks (e.g., eye corners, mouth corners, eyebrow shape), such as the upward angle of mouth corners during smiling or the downward droop of eye corners during sadness;
- Body Movement Features: Extracted from torso posture (e.g., stance, gestures) and motion trajectories (e.g., arm swing amplitude). Intense emotions correlate with exaggerated body movements, while calm emotions correspond to stable postures.

B. MULTIMODAL DATA FUSION TECHNIQUES

Multimodal data fusion enhances the completeness and reliability of emotional features by integrating information from diverse sources. Common fusion levels include data-level, feature-level, and decision-level fusion [15]. This paper adopts a feature-level fusion strategy, integrating features after extraction to balance fusion efficiency and feature expressiveness. The core workflow is: data preprocessing → single-modal feature extraction → feature alignment → multimodal feature fusion. A three-step alignment strategy is adopted to address the temporal discrepancy between audio (10ms/frame) and visual (100ms/frame) data:

Audio downsampling: Audio features are downsampled by averaging within 100ms time windows to match the temporal granularity of visual frames;

Time window matching: Each visual frame is mapped to the corresponding downsampled audio feature window via timestamp synchronization;

Feature interpolation: Linear interpolation is used to supplement missing audio features during performance pauses, ensuring temporal continuity [16].

Alignment Effect: The temporal correlation coefficient between aligned audio and visual features is 0.92. Temporal alignment reduces the model’s emotion change fitting MSE by 12.3%, verifying the effectiveness of the strategy [17].

C. CORE PRINCIPLES OF CONVOLUTIONAL NEURAL NETWORKS

1) 1D-CNN and Audio Feature Extraction

1D-CNN excels at processing temporal data by sliding convolutional kernels to extract local temporal features, making it suitable for both time-domain and frequency-domain feature extraction in audio signals. Its convolution operation is defined as:

$$y_i = \sigma \left(\sum_{k=0}^{K-1} w_k x_{i+k} + b \right) \quad (1)$$

Where x_{i+k} represents the input temporal data segment, w_k denotes the convolution kernel weights, b is the bias term, σ is the activation function (ReLU), and K is the convolution kernel size [18].

TABLE 1. Audio Emotion Feature Classification and Explanation

Feature Type	Specific Feature	Emotional Correlation Description
Acoustic Feature	Mel-Frequency Cepstral Coefficients (MFCC)	Reflects timbre changes, related to emotional delicacy
Acoustic Feature	Fundamental Frequency (F0)	Pitch level, high F0 corresponds to passionate emotion
Acoustic Feature	Short-Time Energy (STE)	Dynamic level, low energy corresponds to sad emotion
Acoustic Feature	Spectral Bandwidth (SBW)	Frequency distribution, wide spectrum corresponds to joyful emotion
Musicological Feature	Beat Per Minute (BPM)	Tempo speed, high BPM corresponds to cheerful emotion
Musicological Feature	Mode (Major/Minor)	Tonality type, major corresponds to bright emotion, minor corresponds to gloomy emotion
Musicological Feature	Note Duration	Note length change, long notes correspond to soothing emotion

2) Applications of 2D-CNN and 3D-CNN in Visual Feature Extraction

2D-CNN extracts spatial features from images using two-dimensional convolutional kernels, suitable for static facial expression feature extraction. 3D-CNN extends 2D-CNN by incorporating temporal dimension convolutions, enabling simultaneous capture of spatial features and temporal variations in visual data, making it suitable for dynamic limb motion feature extraction [19]. The 3D convolution operation formula is as follows:

$$y_{i,j,t} = \sigma \left(\sum_{k=0}^{K-1} w_{a,b,c} x_{i+a,j+b,j+c,t+c} + b \right) \quad (2)$$

where $x_{i+a,j+b,j+c,t+c}$ represents the spatio-temporal segment of the input video frame, $w_{a,b,c}$ denotes the 3D convolution kernel weights, and A, B, C denote the kernel sizes in width, height, and time dimensions, respectively [20].

D. ATTENTION MECHANISM PRINCIPLES

1) Channel Attention Mechanism

The channel attention mechanism enhances key feature channels and suppresses redundant information by calculating importance weights for each feature channel [21]. The core formula is as follows:

$$F_t = \sum_{t=1}^T (\alpha_t f_t) \quad (3)$$

Where F is the input feature map, AvgPool and MaxPool denote global average pooling and global max pooling respectively, MLP is a multilayer perceptron, σ is the Sigmoid activation function, and $M_c(F)$ represents the channel attention weight [22].

III. MULTI-CHANNEL CONVOLUTIONAL ATTENTION MODEL DESIGN

A. OVERALL MODEL ARCHITECTURE

The model adopts a four-layer architecture: “Multimodal Data Input Layer - Multi-Channel Feature Extraction Layer - Hybrid Attention Fusion Layer - Emotional Feature Output Layer.” The input layer receives preprocessed audio and

visual data; the feature extraction layer concurrently extracts multidimensional features through dedicated channels; the hybrid attention fusion layer performs channel-wise and temporal weighting to achieve deep multimodal feature integration; the output layer generates the final emotional feature vector and temporal variation curve [23].

B. CORE MODULE DESIGN

1) Multimodal Data Preprocessing Module

1. Audio Data Preprocessing: Pre-emphasizes (boosts high-frequency components), frames (20ms frame length, 10ms frame shift), applies a Hann window, and performs Fourier transform on captured audio signals to convert them into frequency-domain data. Extracts fundamental features such as MFCC, fundamental frequency, and short-term energy to construct an audio feature matrix (dimensions: $1 \times 200 \times 128$) [24].

2. Visual Data Preprocessing: Extract video frames (10fps), perform face detection and alignment (based on MTCNN algorithm), and extract limb keypoints (based on OpenPose algorithm). Normalize face images (224×224 pixels) and serialize limb motion data to construct a visual feature tensor (dimensions $3 \times 224 \times 224 \times 10$) [25].

2) Multi-Channel Feature Extraction Module

1. Audio Channel (1D-CNN):

Dilated convolution: Dilated rates of 2, 4, 8 are adopted to expand the receptive field without increasing computational complexity, adapting to long-sequence music data; Residual connections: Added between convolutional layers to mitigate gradient vanishing, ensuring effective propagation of long-sequence features [26].

2. Visual Channel (2D-CNN + 3D-CNN):

Facial Expression Subchannel (2D-CNN): 3 convolutional layers (3×3 kernels, 64/128/256 elements), 2 maxpooling layers, 1 fully connected layer, outputting 256-dimensional facial emotion features;

Spatial attention modules: Embedded in 2D-CNN to focus on emotion-related regions (mouth, eyes, body key points); Asymmetric 3D kernels: Use $3 \times 3 \times 5$ kernels to enhance capture of short-term dynamic changes in performances [27].

3) Hybrid Attention Fusion Module

1. Channel Attention Submodule: Inputs audio and visual feature vectors. Captures channel feature statistics via global average pooling and max pooling. Generates channel attention weights (dimensions 512×1) through an MLP. Applies weighted fusion to audio and visual features:

$$F'_a = F_a \odot M_c(F_a), \quad F'_v = F_v \odot M_c(F_v) \quad (4)$$

Where F_a and F_v represent original audio and visual features, $M_c(\cdot)$ denotes channel attention weights, F'_a , F'_v denote channel-weighted features [28]. $\odot$ denotes element-wise multiplication.

2. Temporal Attention Submodule: Arrange the channel-weighted features by time step to construct a temporal feature sequence (dimension $T \times 1024$, where T is the number of time steps). Compute weights for each time step via temporal attention and perform weighted fusion on the temporal sequence:

$$F_{tv} = \sum_{t=1}^T \alpha_t [F'_{a,t} \parallel F'_{v,t}] \quad (5)$$

where α_t denotes the temporal attention weight at time step t , $\parallel$ indicates feature concatenation, and F_{tv} represents the temporally weighted multimodal fusion feature [29].

3. Feature Fusion Layer: An adaptive weighting strategy dynamically adjusts the fusion weights between audio and visual features:

$$F_{\text{final}} = \beta F'_a + (1 - \beta) F'_v \quad (6)$$

where beta and gamma are trainable weight coefficients (optimized to 0.4 and 0.3 via backpropagation in experiments), and F_{final} is the final emotion feature vector.

Long-Sequence Processing Strategy: For performances up to 15 minutes, a “5-second sliding window + temporal attention fusion” strategy is adopted. Each window’s features are extracted independently, and temporal attention weights integrate window-level features to balance local detail capture and global emotional tracking. This strategy improves the accuracy of complex long-sequence data processing by 8.7% [30].

4) Emotion Feature Output Module

Outputs two forms of emotion features: A fixed-dimensional (1024-dimensional) emotion feature vector for emotion classification tasks; A temporal emotion intensity sequence, generated by quantifying emotion intensity at each time step to produce an emotion evolution curve, visually reflecting the dynamic evolution of emotions during performances.

IV. EXPERIMENTAL VALIDATION AND DESIGN

A. DATASET CONSTRUCTION AND DESCRIPTION

1) Self-Built University Music Performance Emotion Dataset (UMPED)

1. Data Collection: Performance videos and audio recordings were collected from vocal (bel canto, ethnic, pop), instrumental (piano, violin, guzheng) students at a university music conservatory. A total of 120 performances were recorded, ranging from 2 to 15 minutes per session, with a total duration of 860 minutes. This study was approved by the Ethics Committee of Hefei Normal University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

2. Data Annotation: A team comprising 3 music education experts and 2 researchers in affective computing annotated emotional categories (joy, sadness, excitement, relaxation, calmness) and emotional intensity (1–5 points, higher scores indicating stronger emotion). Annotation consistency was validated using Cohen’s Kappa coefficient (Kappa=0.87, indicating high agreement).

3. Sample Segmentation Standards: Audio samples are segmented by “musical phrases” (3–5 seconds per phrase, with 4,300 complete phrases in 120 performances), ensuring alignment with the emotional expression logic of music. Visual samples are segmented by “expression and action segments”—one sample is extracted every 2 frames based on keyframe detection, resulting in 25,800 valid segments across all performances.

4. Temporal Continuity Assurance: Each audio sample is labeled with performance ID, phrase number, and start-end timestamp. Each visual sample is assigned a corresponding audio sample association ID. During model training, temporal synchronization of cross-modal data is achieved through timestamp alignment, avoiding fragmentation of temporal information.

2) Public Dataset (MusicNet)

The publicly available MusicNet emotional music dataset was used for supplementary validation. This dataset comprises 500 classical music excerpts annotated with four emotional categories—joy, sadness, anger, and calm—spanning diverse instruments and performance styles, enabling verification of the model’s generalization capability.

B. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

1) Experimental Environment

Hardware: Intel Core i9-13900K CPU, NVIDIA RTX 4090 GPU (24GB), 64GB RAM; Software: Python 3.9, PyTorch 2.0, OpenCV 4.8, Librosa 0.10 (audio processing library), MTCNN (face detection library).

2) Experimental Parameter Settings

Model training parameters: Batch Size = 32, Learning Rate = 0.001, Epoch = 100, Weight Decay = 0.0001, Dropout Rate = 0.3; AdamW optimizer used; cross-entropy loss (for

classification tasks) and mean squared error loss (for time series fitting tasks) employed as loss functions. Overfitting Control Measures:

1. Balanced sample sampling: Sufficient samples are added for each emotion-feature combination to avoid bias;
2. Dropout layer: A dropout rate of 0.3 is adopted in the fully connected layers;
3. L2 regularization: Weight decay coefficient $\lambda = 0.0001$ to limit parameter scale.

C. COMPARATIVE MODEL SELECTION

Five mainstream sentiment feature extraction models were selected for comparison:

1. Single-Modality Models: Audio sentiment extraction based on 1D-CNN, visual sentiment extraction based on 2D-CNN;
2. Traditional Multimodal Fusion Models: CNN + Feature Concatenation (Concat), CNN + Weighted Sum;
3. Single-Attention Models: CNN + Channel Attention (CA), CNN + Temporal Attention (TA);
4. Classic Deep Learning Models: LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit);
5. Recent Models: Transformer-based Sentiment Extraction Models.

D. EXPERIMENTAL DESIGN

1. Sentiment Classification Experiment: Compare sentiment classification accuracy across models using the UMPED and MusicNet datasets;
2. Temporal Sentiment Change Fitting Experiment: Select 10 performance segments with distinct emotional shifts and calculate the Mean Squared Error (MSE) for temporal fitting across models;
3. Ablation experiments: Validate the effectiveness of multi-channel architecture, channel attention, and temporal attention modules;
4. Robustness experiments: Test sentiment extraction performance under noisy audio and blurred visual data conditions.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. SENTIMENT CLASSIFICATION ACCURACY ANALYSIS

Sentiment classification accuracy comparisons across models on the UMPED and MusicNet datasets are shown in Table 2.

TABLE 2. Comparison of Sentiment Classification Accuracy

Model	UMPED Dataset	MusicNet Dataset	Average Accuracy
1D-CNN (Audio Single Modality)	76.5	78.3	77.4
2D-CNN (Visual Single Modality)	74.2	75.1	74.6
CNN+Concat	82.7	83.5	83.1
CNN+Weight Sum	83.4	84.2	83.8
CNN+CA	85.6	86.3	85.9
CNN+TA	86.1	87.2	86.6
LSTM	84.3	85.5	84.9
Transformer-based	88.7	89.5	89.1
Proposed Model	94.3	93.8	94.0

As shown in Table 2, our model achieves significantly higher classification accuracy than the comparison models on both datasets, with an average accuracy of 94.0%. This represents an improvement of 16.6%–19.4% over single-modal models, 10.2%–11.9% over traditional multimodal fusion models, and 7.4%–8.1% over single-attention models. This superiority stems from the proposed model's dual-channel parallel extraction of audio and visual features, combined with a hybrid attention mechanism that amplifies key information, thereby fully leveraging the complementary nature and temporal dynamics inherent in multimodal data. Reasons for High Accuracy:

High-quality annotation: The UMPED dataset has a Kappa coefficient of 0.87, covering 5 fine-grained emotions with strong sample specificity;

Multimodal complementarity: The multi-channel architecture fully integrates audio and visual features, providing more comprehensive information than single-modal models;

Targeted attention mechanism: The hybrid attention mechanism amplifies key emotional features and suppresses redundant noise, enhancing feature discriminability. **Risk Verification:**

Data isolation: Training and test sets use performances from different batches of students, with no overlap. Data preprocessing and annotation are completed by separate teams to avoid data leakage;

Generalization validation: The model achieves 93.8% accuracy on the public MusicNet dataset, only 0.5% lower than on the UMPED dataset, confirming no overfitting.

B. ANALYSIS OF TEMPORAL SENTIMENT CHANGE FITTING

The mean squared error (MSE) for temporal sentiment change fitting across models is compared in Table 3.

TABLE 3. Comparison of MSE in Time-Series Emotion Change Fitting

Model	Joy-Energetic	Sad-Soothing	Calm-Sad	Average MSE
1D-CNN (Audio)	0.087	0.092	0.076	0.085
2D-CNN (Visual)	0.093	0.098	0.082	0.091
CNN+Concat	0.065	0.071	0.062	0.066
CNN+CA	0.058	0.063	0.055	0.059
CNN+TA	0.042	0.047	0.039	0.043
Transformer-based	0.038	0.042	0.035	0.038
Proposed Model	0.031	0.034	0.029	0.032

The average temporal fitting MSE of this model is only 0.032, representing a reduction of 62.4%–64.8% compared to single-modal models, 51.5% compared to traditional multimodal models, and 25.6% compared to single temporal attention models. This demonstrates that the hybrid attention mechanism effectively captures the dynamic evolution of sentiment over time, particularly during sentiment transition periods, where the model accurately tracks fluctuations in sentiment intensity.

C. ABLATION STUDY ANALYSIS

The ablation study results are presented in Table 4. By progressively removing the multi-channel structure, channel

attention, and temporal attention modules, we observe the corresponding changes in model performance.

TABLE 4. Ablation Study Results

Model Configuration	Emotional Classification Accuracy	Performance Drop Rate
Proposed Complete Model	94.3	-
Remove Multi-Channel Structure (Single-Channel Extraction)	87.6	6.7%
Remove Channel Attention	89.2	5.1%
Remove Temporal Attention	90.5	3.8%
Keep Only Multi-Channel + No Attention	86.3	8.0%

Ablation experiments demonstrate that the multi-channel architecture, channel attention, and temporal attention all significantly contribute to model performance. Among these, the multi-channel architecture exerts the strongest influence, as its parallel extraction of multimodal features forms the foundation for accurately capturing sentiment characteristics. Channel attention and temporal attention respectively enhance the capture of key features and dynamic changes, with their synergistic interaction further elevating model performance.

D. ROBUSTNESS ANALYSIS

Under scenarios involving 10dB and 20dB noise added to audio, and visual image resolution reduced by 50% and 70%, the robustness test results for the proposed model and comparison models are shown in Table 5.

TABLE 5. Robustness Test Results

Experimental Scenario	Proposed Model	Transformer-based CNN+TA	1D-CNN (Audio)	
Original Data	94.3	88.7	86.1	76.5
Audio with 10dB Noise	91.2	84.3	81.5	68.2
Audio with 20dB Noise	87.5	79.6	76.8	61.3
Visual Resolution Reduced by 50%	92.6	86.5	83.4	-
Visual Resolution Reduced by 70%	89.8	82.1	79.2	-

As shown in Table 5, under scenarios of deteriorating data quality, our model exhibits the smallest accuracy decline (original $\rightarrow$ 20dB noise: 6.8% decrease; Original $\rightarrow$ 70% resolution reduction: 4.5% decrease), significantly outperforming the comparison models. This advantage stems from the complementary nature of multimodal data and the attention mechanism's ability to focus on key features. Even when a single modality is disrupted, the model achieves precise emotion extraction by leveraging the other modality and weighting critical features.

E. CASE STUDY ANALYSIS

A 3-minute piano recital clip from a university was selected, featuring a complete emotional progression from "Calm $\rightarrow$ Joy $\rightarrow$ Excitement $\rightarrow$ Serenity." The prediction curve closely aligns with the actual curve, with response delays at emotional transition points (60 seconds: Calm $\rightarrow$ Joy; 120 seconds: joy $\rightarrow$ passion; 160 seconds: passion $\rightarrow$ calm),

with a response delay of less than 0.5 seconds, enabling precise capture of the dynamic emotional evolution process. Potential risks of this research include:

1. Sample source limitation: Samples are collected from a single university, which may have geographical and teaching mode biases;
2. Annotation subjectivity: Manual annotation of fine-grained emotions has inherent subjective differences, which may slightly affect model training.
3. Future improvements will expand sample sources to multiple universities and increase the number of annotators to enhance result reliability.

VI. CONCLUSIONS AND OUTLOOK

A. RESEARCH FINDINGS

Addressing the precision and dynamic capture requirements for emotional feature extraction in university music performances, this study designed an extraction method based on a multi-channel convolutional attention mechanism. Key achievements include:

1. Established the University Music Performance Multi-modal Emotion Dataset (UMPED), encompassing vocal and instrumental performances, providing high-quality data support for emotion feature extraction research;
2. Proposed a multi-channel convolutional feature extraction network, designing dedicated channels for the heterogeneous characteristics of audio and visual modalities to achieve parallel and efficient extraction of multidimensional emotional features;
3. Integrated channel attention and temporal attention to construct a hybrid attention mechanism, enabling adaptive weighting of key emotional features and precise capture of dynamic changes;
4. Experimental validation demonstrated an emotion classification accuracy of 94.3% and a temporal emotion change fitting MSE as low as 0.032, representing significant improvements over traditional models. The method exhibits strong robustness under noisy and ambiguous data conditions. This approach effectively addresses the limitations of traditional methods in insufficient multimodal fusion and inadequate temporal dynamic capture. It accurately extracts emotional feature changes in university music performances, providing an objective and quantifiable technical means for music education evaluation and performance quality analysis.

B. FUTURE PROSPECTS

Future research can be deepened in three aspects:

1. Expand modal dimensions by integrating performers' physiological data (e.g., heart rate, respiration rate) with musical score information to further enhance the comprehensiveness of emotional features;
2. Optimize attention mechanism design by introducing cross-modal attention to strengthen the extraction of associative information between audio and visual modalities;
3. Build Practical Application Platform: Develop an integrated platform for data collection, feature extraction,

and evaluation analysis. The platform will support real-time emotional feature extraction during performances and provide quantitative feedback for music education.

AUTHOR CONTRIBUTIONS

Qin Kang designed, collected and analyzed the data, and drafted the manuscript. Qin Kang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Qin Kang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, 2022 Horizontal Project: "Reform of Piano Education and Training System and Collaborative Education Construction Under the 'Double Reduction' Policy" (HXXM2022093).

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Hefei Normal University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] W. Yang, Y. Liu, Y. Chen, Y. Cui, C. Guo, G. Shen, et al., "HSMNet: A multi-resolution grayscale image steganalysis method based on hybrid dilated convolution and self-attention multi-channel network," *Information Sciences*, vol. 728, 122824, 2026, doi: 10.1016/j.ins.2025.122824.
- [2] R. Li, G. Zhang, Y. Niu, K. Rong, W. Liu, and H. Hong, "A multi-channel multi-scale spatiotemporal convolutional cross-attention fusion network for bearing fault diagnosis," *Sensors (Basel, Switzerland)*, vol. 25, no. 18, 5923, 2025, doi: 10.3390/s25185923.
- [3] H. Mulam, V. Rameshkumaar, M. Bachute, and E. Muniyandy, "Dual attention holographic convolutional neural network based ictal and interictal states of automatic seizure detection using multi-channel scalp EEG with Humboldt squid optimization algorithm," *Smart Health*, vol. 37, 100599, 2025, doi: 10.1016/j.smhl.2025.100599.
- [4] P. Zhou, Z. Qian, D. Fu, and Y. Zhang, "Plunger pump fault diagnosis method based on wavelet convolution and multi-head self-attention mechanism for multi-channel feature fusion," *Journal of Computer and Communications*, vol. 13, no. 7, pp. 345-356, 2025, doi: 10.4236/jcc.2025.137018.
- [5] M. Yang, R. Jiang, B. Wang, G. Fang, Y. Jia, and F. Fan, "Multi-channel attention mechanism graph convolutional network considering cumulative effect and temporal causality for day-ahead wind power prediction," *Energy*, vol. 332, 137023, 2025, doi: 10.1016/j.energy.2025.137023.
- [6] Y. Lin, S. Pan, J. Yu, Y. Hong, F. Wang, L. Zheng, et al., "MDCA-DETR: DETR with multi-channel deformable convolution and coordinate attention for mini-LED wafer surface defects detection," *Optics and Lasers in Engineering*, vol. 193, 109082, 2025, doi: 10.1016/j.optlaseng.2025.109082.
- [7] X. Wu, Y. Fu, Q. Wu, Z. Wen, H. Dong, and D. Zhang, "Gearbox fault diagnosis based on multi-channel orthogonal attention network," *Signal, Image and Video Processing*, vol. 19, no. 8, 651, 2025, doi: 10.1007/s11760-025-04132-5.
- [8] L. Luo, X. Fan, M. Wang, Z. Liu, and Y. Zhang, "Two-stage deep complex gated convolution attention network for multi-channel noise suppression," *Signal, Image and Video Processing*, vol. 19, no. 7, 512, 2025, doi: 10.1007/s11760-025-04158-9.
- [9] M. V. Divya, "Multi-channel mantis search spatial attention convolutional neural network based hyperspectral image change detection," *Multimedia Tools and Applications*, vol. 84, no. 33, pp. 1-22, 2025, doi: 10.1007/s11042-025-20810-5.
- [10] R. Zhao, C. Zhang, and D. Xue, "A multi-scale multi-channel CNN introducing a channel-spatial attention mechanism hyperspectral remote sensing image classification method," *European Journal of Remote Sensing*, vol. 57, no. 1, 2353290, 2024, doi: 10.1080/22797254.2024.2353290.
- [11] A. G. Kakisim and Z. Turgut, "Multi-channel convolutional neural network with attention mechanism using dualband WiFi signals for indoor positioning systems in smart buildings," *Internet of Things*, vol. 29, 101435, 2025, doi: 10.1016/j.iot.2024.101435.
- [12] Z. Qu, L. Zhou, X. Yin, and T. Lu, "A crack detection network with multi-channel attention and enhanced information interaction," *Signal, Image and Video Processing*, vol. 19, no. 1, 37, 2024, doi: 10.1007/s11760-024-03581-8.
- [13] L. Shi, W. Liu, Y. Wu, C. Dai, Z. Ji, and I. Ganchev, "Knowledge graph embedding using a multi-channel interactive convolutional neural network with triple attention," *Mathematics*, vol. 12, no. 18, 2821, 2024, doi: 10.3390/math12182821.
- [14] J. Liu, Y. Huang, D. Wu, Y. Yang, Y. Chen, L. Chen, et al., "Multi-channel multi-scale convolution attention variational autoencoder (MCA-VAE): An interpretable anomaly detection algorithm based on variational autoencoder," *Sensors*, vol. 24, no. 16, 5316, 2024, doi: 10.3390/s24165316.
- [15] S. Liu, H. Zhang, A. Zhang, J. Liu, and Y. Liu, "Subsurface temperature estimation of mesoscale eddies in the Northwest Pacific Ocean from satellite observations using a residual multi-channel attention convolution network," *Frontiers in Marine Science*, vol. 11, 1397109, 2024, doi: 10.3389/fmars.2024.1397109.
- [16] Y. Wang, Y. Zhou, W. Lu, Q. Wu, Q. Li, and R. Zhang, "AC-CfC: An attention-based convolutional closed-form continuous-time neural network for raw multi-channel EEG-based emotion recognition," *Biomedical Signal Processing and Control*, vol. 94, 106249, 2024, doi: 10.1016/j.bspc.2024.106249.
- [17] M. Zhu and L. Wen, "A multi-channel convolutional neural network based on attention mechanism fusion for facial expression recognition," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, 2, 2024, doi: 10.2478/amns.2023.1.00084.
- [18] X. Yang, J. Zhu, and J. Huang, "A multi-channel CNN fault diagnosis method based on squeeze-and-convolution attention for rotating machinery: Papers," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 18, no. 8, JAMDSM0097, 2024, doi: 10.1299/jamds.2024jamds0097.
- [19] J. He, H. Yang, S. Zhou, J. Chen, and M. Chen, "A dual-attention-mechanism multi-channel convolutional LSTM for short-term wind speed prediction," *Atmosphere*, vol. 14, no. 1, 71, 2022, doi: 10.3390/atmos14010071.
- [20] V. Mayya, S. Kamath, G. S. Krishnan, and T. Gangavarapu, "Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries," *Future Generation Computer Systems*, vol. 118, pp. 374-391, 2021, doi: 10.1016/j.future.2021.01.013.
- [21] Y. Geng, L. Wang, J. Hu, S. Luo, and Q. Zhou, "A dual category-channel attention convolutional neural network fault diagnosis method for imbalanced multi-sensor data of spacecraft," *Aerospace Science and Technology*, vol. 170, 111593, 2026, doi: 10.1016/j.ast.2025.111593.
- [22] J. Liao and L. Wang, "MSDANet: A multiscale dual-channel spatial attention network with depthwise separable convolution for hyperspectral image classification," *Canadian Journal of Remote Sensing*, vol. 51, no. 1, 2551533, 2025, doi: 10.1080/07038992.2025.2551533.
- [23] J. Yin, Z. Huang, Q. Chen, W. Huang, L. Sun, Q. Wu, et al., "Cascaded local-nonlocal pansharpening with adaptive channel-kernel convolution and multi-scale large-kernel attention," *Remote Sensing*, vol. 18, no. 1, 97, 2025, doi: 10.3390/rs18010097.
- [24] Y. Tang and Z. Luo, "A dual channel fusion method for wind power prediction based on graph attention network and wavelet convolutional

- neural network,” *Electric Power Systems Research*, vol. 254, 112659, 2026, doi: 10.1016/j.epsr.2025.112659.
- [25] M. Zolfaghari, A. M. Saniee, and H. Sajedi, “Channel-spatial attention modules in convolutional neural networks for image classification,” *Scientific Reports*, vol. 16, no. 1, 2348, 2025, doi: 10.1038/s41598-025-32055-1.
- [26] J. A. Ramirez-Quintana, J. J. Muñoz-Pacheco, G. Ramirez-Alonso, J. A. Medrano-Hermosillo, and A. D. Corral-Saenz, “Lightweight convolutional neural network with efficient channel attention mechanism for real-time facial emotion recognition in embedded systems,” *Sensors*, vol. 25, no. 23, 7264, 2025, doi: 10.3390/s25237264.
- [27] A. Salam, M. Abrar, R. W. Anwer, F. Amin, F. Ullah, I. de la Torre, et al., “Channel-attention DenseNet with dilated convolutions for MRI brain tumor classification,” *Computer Modeling in Engineering & Sciences*, vol. 145, no. 2, pp. 2457-2479, 2025, doi: 10.32604/cmescs.2025.072765.
- [28] Y. Lin, M. Zhao, and B. Su, “Lcpa-msdce UNet: A unet variant integrating lightweight channel-pixel attention and multi-scale dilated convolution enhancement modules for medical image segmentation,” *The Journal of Supercomputing*, vol. 81, no. 17, 1581, 2025, doi: 10.1007/s11227-025-08069-x.
- [29] Y. Zhang, B. Qi, H. Yu, Y. Shi, and Z. Huang, “MSDC-CCA: DWT-based infrared image destripper with multi-scale dilated convolution and coordinate-channel attention,” *Infrared Physics and Technology*, vol. 152, 106242, 2026, doi: 10.1016/j.infrared.2025.106242.
- [30] S. Luo, X. Gong, W. Du, L. Wang, K. Feng, and Y. Qian, “Small sample fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighted temporal convolutional network (DCK-CAM-TCN),” *Journal of Vibroengineering*, vol. 27, no. 7, pp. 1278-1295, 2025, doi: 10.21595/jve.2025.25034.