

Article

DC-MEPV: Dual-Channel Assisted Music Emotion Perception and Visualization in Acousto-Optic Synergistic Intelligent Cockpits

Wei Shen ¹, Xingang Mou ^{1,2} , Songqing Le ¹, Zhixing Zong ¹ and Jiaji Li ^{1,2,*} 

¹ School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China; pianle922@gmail.com (W.S.); sunnymou@whut.edu.cn (X.M.); lesongqing_edu@163.com (S.L.); zongzhixing_edu@126.com (Z.Z.)

² Key Laboratory of Advanced Manufacturing Technology for High Performance Parts, Ministry of Education, Wuhan 430070, China

* Correspondence: jiajili@whut.edu.cn

Abstract

We propose a Dual-Channel assisted Music Emotion Perception and Visualization (DC-MEPV) framework designed for ambient lighting in intelligent vehicle cockpits, addressing the increasing demand for advanced human–machine interaction in the automotive industry. This framework consists of three main components: the Multi-Scale Feature Extraction Block (MSFEB), the Global Sequence Modeling Block (GSMB), and the Emotional Color Visualization Algorithm (ECV-Algo). The MSFEB extracts valence and arousal (V-A) features from dual channels at multiple temporal scales, with each channel employing a hybrid neural network architecture to capture multi-scale emotional representations. The GSMB integrates positional encoding, bidirectional long short-term memory (BiLSTM) networks, and multi-head self-attention mechanisms to dynamically model global emotional sequences. The ECV algorithm utilizes personalized emotion–color association rules to achieve expressive emotion-driven lighting visualization based on a continuous mapping from emotion space to color space. We conducted comprehensive comparison and ablation experiments to evaluate the model’s emotion perception performance, and designed three metrics to evaluate the quality of the generated visualizations. The model outperformed other networks in both comparative and ablation experiments. Additionally, the generated lights demonstrated strong performance in terms of CIEDE2000 variation rates, unique color ratios, and joint histogram entropy. DC-MEPV achieved excellent performance in emotion perception and visualizations on the DEAM and PMemo datasets.

Keywords: music emotion perception and visualization; multi-scale feature extraction; music-to-emotional-color mapping; intelligent cockpit



Academic Editor: Douglas O’Shaughnessy

Received: 26 February 2026

Revised: 3 April 2026

Accepted: 7 April 2026

Published: 13 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the rapid advancement of automotive technologies, intelligent systems have significantly improved driving convenience, comfort, and in-car entertainment [1]. However, their potential to enhance driving safety remains largely unexplored. Extensive evidence suggests that a driver’s emotional state has a significant impact on driving behavior and road safety. Negative emotions such as anger or anxiety often lead to dangerous driving behaviors, while positive emotions help maintain focus and make more informed decisions [2–4]. Music, as a powerful medium for emotion regulation, exhibits strong

synesthetic associations with color. The real-time mapping of musical emotions to ambient lighting colors for the joint regulation of drivers' emotional states represents a promising development direction for intelligent vehicle cockpits. Such systems can promote safer driving by actively regulating driver emotions, but this remains an important and under-explored area of research [5]. Accurate extraction and quantification of musical emotion form the foundation of music-to-color synesthesia.

Early music emotion recognition (MER) systems relied on handcrafted acoustic features, such as mel-frequency cepstral coefficients (MFCCs) [6,7] and spectrograms [8], combined with traditional machine learning methods including Gaussian Mixture Models (GMMs) [9,10], Support Vector Regression (SVR) [11], and Bayesian classifiers [12]. However, manual feature engineering is limited by subjective bias, often failing to capture nuanced affective information [13,14] and introducing redundant variability that compromises model robustness [15–17]. In recent years, deep learning has emerged as the dominant paradigm in MER by enabling end-to-end feature learning [18]. Current approaches can be broadly classified into audio-based, lyric-based, and multimodal frameworks. Among audio-based methods, Li et al. [19] employed a BiLSTM architecture to capture long-range temporal dependencies for dynamic emotion tracking. Yang et al. [20] proposed a CNN-LSTM hybrid to alleviate gradient degradation in long sequences; other notable models include Inception-GRU with residual connections [21] and the Bidirectional Convolutional Recurrent Sparse Network (BCRSN) [22]. Lyric-based methods, such as those by Malheiro et al. [23] and Angelica et al. [24], exploit textual semantics using regression or LSTM-GRU ensembles. Multimodal approaches, exemplified by Zhao et al. [25], fuse lyrics, metadata, and artist information via pretrained BERT embeddings and cross-modal attention, achieving substantial improvements in recognition accuracy. Once musical emotion is quantified as a temporal sequence; the next step involves mapping these emotional states to colors for ambient lighting. Existing mapping strategies include physics-based methods [26], Kansei engineering [27], subjective assessments [28], and direct associations between color samples and emotional values [29–31]. However, these approaches heavily depend on subjective evaluations and are therefore prone to high inter-individual variability [28,31]. In practical lighting implementations, current systems often rely on subjective surveys [32], IRI color models combined with Thayer's emotional framework for LED control [33], or automated adjustments based on musical segments [34]. Nevertheless, most existing mappings are discrete and static, failing to reflect continuous emotional dynamics. Related studies have also demonstrated that colored lighting can actively influence mood [35,36].

Despite these advances in music-driven emotional color transformation, several key limitations remain. First, most MER models extract the V-A dimension through a single shared path, which fails to explain the psychological independence of the valence (pleasure) and arousal (activation) dimensions, since valence is more related to spectral content, while arousal is related to time dynamics. Second, current emotion-to-color mappings lack personalization, temporal continuity, and dynamic responsiveness. To address these gaps, this paper proposes a dual-channel-assisted framework for music emotion perception and visualization. The framework incorporates multi-scale feature extraction, BiLSTM, and multi-head self-attention mechanisms to achieve more accurate and robust emotional representation. Additionally, the proposed ECV-Algo is introduced to enable continuous, personalized, and dynamic mapping from emotional states to color spaces, thereby enhancing acousto-optic synergy and opening new possibilities for intelligent cockpit applications. Generally, the novelty of DC-MEPV stems from its dual-channel design that decouples V-A feature extraction for specialized multi-scale modeling and the ECV-Algo's personalized continuous mapping, which surpasses the limitations of single-pathway MER and static mappings in prior works.

2. Methods

The proposed DC-MEPV is a dual-channel-assisted music-to-emotional-color transformation framework designed to enable dynamic acousto-optic synergy between music emotion and ambient lighting in intelligent vehicle cockpits. The framework seamlessly integrates a dual-channel architecture, multi-scale feature extraction, global temporal sequence modeling, and personalized emotion-to-color mapping to significantly enhance drivers' emotion regulation capabilities [37]. The overall system architecture is illustrated in Figure 1. The input audio signal is first preprocessed using Fourier transform to produce a Mel spectrogram, which serves as the input to the subsequent modules. This spectrogram is processed in parallel by two MSFEBs with different receptive fields, which allows for the independent capture of emotional features along the V-A dimensions. The extracted sentiment features are then input into GSMB to predict consecutive V-A sequences. Finally, the proposed ECV-Algo maps the predicted V-A values to the HSV color space. The algorithm incorporates personalized emotion–color association rules and intensity-guided dynamic interpolation to generate smooth, temporally coherent ambient lighting sequences that synchronize with the evolving musical emotion.

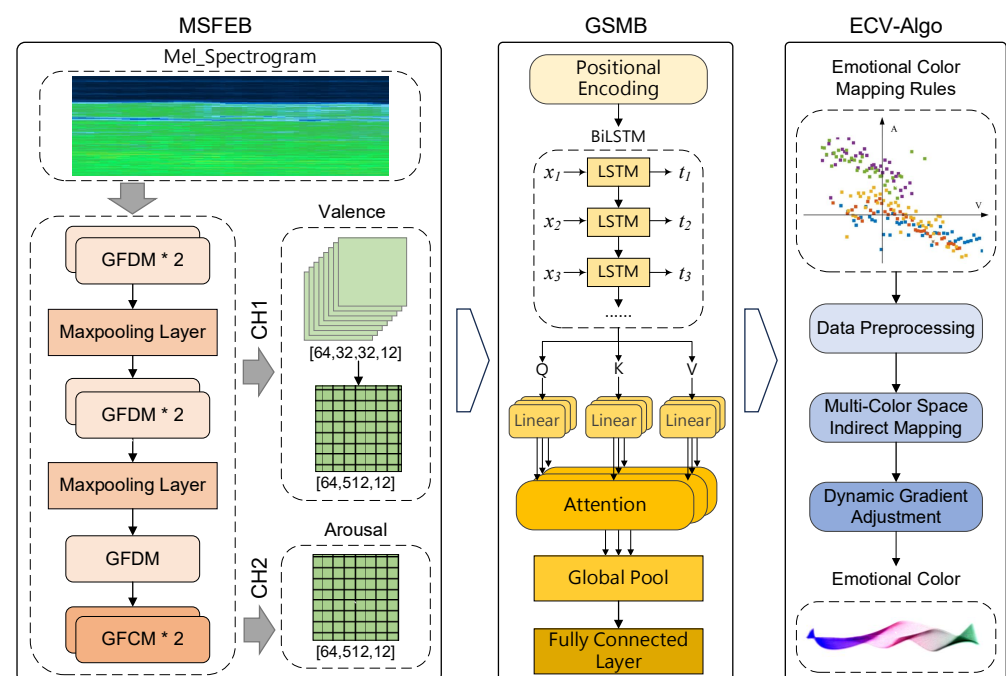


Figure 1. Overall architecture of the proposed DC-MEPV framework. The diagram depicts the complete pipeline from raw audio input to final RGB color sequence output. The left section illustrates the dual-channel feature extraction pathways (valence channel on the upper and arousal channel on the lower), each followed by an independent Global Sequence Modeling Block. The right section shows the ECV-Algo, which comprises data preprocessing, color mapping, and dynamic gradient adjustment for smooth transitions.

2.1. Multi-Scale Feature Extraction Block

The Multi-Scale Feature Extraction Block (MSFEB) effectively captures V-A emotional features from music audio. The two channels utilize distinct MSFEB configurations with tailored convolutional kernel sizes to optimize feature extraction for the respective valence and arousal dimensions. Each MSFEB consists of a stack of Grouped Feature Extraction Modules (GFEMs) interleaved with max-pooling layers. This design allows the model to learn hierarchical features directly from the Mel spectrogram. A Feature Fusion Module (FFM) is added to the valence channel to reduce its output dimension to be the same as

that of the arousal channel, so that the input of the subsequent block remains consistent in dimension. The GFEM forms the core component of the MSFEB. As shown in Figure 2, the network processes the input Mel spectrogram through four parallel branches, and each convolutional kernel has different scales to capture multi-resolution features. The branch outputs are concatenated and fused across groups to produce an enriched representation. This design accounts for the distinct characteristics of the valence and arousal dimensions. The valence dimension, which reflects pleasantness versus sadness, is strongly correlated with spatial spectral patterns. Thus, the valence channel employs 2D convolutions with kernel sizes of 1×1 , 3×3 , and 5×5 to extract local and global structures. In contrast, the arousal dimension, which reflects excitement versus calmness, is more closely tied to temporal dynamics. Accordingly, the arousal channel uses 1D convolutions with kernel lengths of 1, 3, 5, and 7 to efficiently model sequential evolution [38]. By tailoring kernel configurations to each dimension and leveraging grouped convolutions, the GFEM effectively captures diverse rhythmic and timbral expressions while reducing computational overhead.

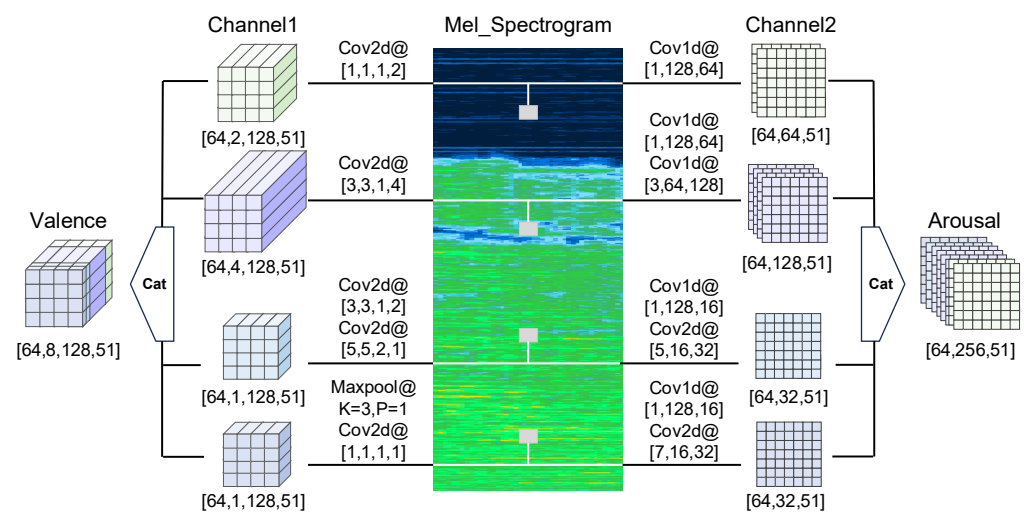


Figure 2. Architecture of the Grouped Feature Extraction Module (GFEM). The diagram illustrates four parallel branches with different kernel sizes, followed by group concatenation and fusion. The valence channel employs 2D convolutions (kernel sizes 1×1 , 3×3 , 5×5), whereas the arousal channel uses 1D convolutions (kernel lengths 1, 3, 5, 7), emphasizing dimensional specialization.

GFEM can be configured in two variants. In the forward direction, the network is called Grouped Feature Dilation Module (GFDM), whereas in the reverse direction, it is called Grouped Feature Compression Module (GFCM). Each MSFEB consists of two GFDM layers for feature expansion, followed by a max-pooling layer, another two GFDM layers, a second max-pooling layer, one additional GFDM layer, and finally two GFCM layers for feature compression. The initial GFDM layers enhance representational capacity through multi-scale grouped convolutions to generate rich spatiotemporal emotional features. Subsequent max-pooling operations downsample the feature maps along both time and frequency axes to reduce computational complexity. The intermediate GFDM layers extract higher-level abstractions to capture more intricate affective patterns. After further downsampling, the final GFDM layer refines the representation, and the two GFCM layers progressively reduce channel depth to produce compact yet discriminative features suitable for downstream temporal modeling. To ensure dimensional consistency, we integrate the Feature Fusion Module (FFM) into the valence branch. The FFM flattens the 2D feature map and divides it into four subsequences for subsequent processing. Each subsequence is transposed and reorganized channel-wise to form four feature blocks. A single-head self-attention mechanism then computes weights for each block, generating

compressed outputs that emphasize informative elements while suppressing redundancy. The four compressed blocks are concatenated along the sequence dimension, and a final 1D max-pooling layer halves the sequence length to match the arousal channel output. The FFM employs a single-head self-attention mechanism with a hidden dimension of 512 to dynamically compute importance weights for each feature block, effectively resolving dimensional inconsistencies while preserving the most discriminative emotional information. Compared to conventional feature concatenation approaches in hybrid models [20,39], the FFM effectively resolves dimensional inconsistencies and mitigates redundancy after cross-dimensional processing.

2.2. Global Sequence Modeling Block

The Global Sequence Modeling Block (GSMB) serves as the core module for producing accurate continuous V-A predictions. It integrates positional encoding, bidirectional long short-term memory (BiLSTM), multi-head self-attention (MA), 1D global average pooling, and a fully connected layer to enable dynamic modeling of spatiotemporal emotional features. The input sequence is first positionally encoded to inject relative positional information in order to compensate for the permutation invariance of the attention mechanism. The BiLSTM layer then captures bidirectional temporal dependencies, while the multi-head self-attention layer extracts global inter-feature correlations. Finally, global average pooling aggregates the sequence into a compact representation, which the fully connected layer regresses into continuous V-A scores. This architecture mitigates gradient vanishing in long sequences, enhances global dependency modeling, and strengthens dimensional decoupling [40,41]. To preserve dimensional independence while reducing computational overhead, separate GSMB instances are deployed for each channel with shared parameters.

The positional encoding layer plays a critical role in the GSMB by injecting positional information into the input sequence. This enables subsequent BiLSTM and multi-head self-attention mechanisms to discern the relative order of elements—a factor particularly important in music emotion recognition, where temporal attributes such as rhythmic transitions and emotional peaks directly affect V-A prediction accuracy [42]. Although conventional recurrent models inherently capture sequential order, multi-head self-attention is permutation-invariant and risks treating the sequence as a bag-of-features without positional cues [43]. This study adopts fixed sinusoidal positional encoding, defined as

$$PE(p, k) = \begin{cases} \sin[p / (1E + 8n/d)] & \text{if } k = 2n, \\ \cos[p / (1E + 8n/d)] & \text{if } k = 2n + 1, \end{cases} \quad (1)$$

where p is the position, k is the dimension index, and d is the model dimension. The positional vectors are added element-wise to the input sequence before further processing.

The BiLSTM layer forms the primary temporal modeling component of the GSMB. It processes the position-encoded sequence using a bidirectional architecture comprising stacked LSTM units, each equipped with forget, input, and output gates to regulate long-term memory. Compared to unidirectional LSTMs [19], BiLSTM significantly improves dynamic prediction accuracy, especially for the arousal dimension in music with varying tempo. Multi-head self-attention complements the BiLSTM by effectively modeling long-range dependencies that recurrent layers struggle to capture. This mechanism projects the input into multiple subspaces (“heads”), computes scaled dot-product attention independently in each head, concatenates the outputs, and applies a linear transformation. The multi-head self-attention employs 8 heads with scaled dot-product attention and a dropout rate of 0.1, allowing the model to jointly capture diverse global inter-feature correlations in both valence and arousal sequences. The parallel heads allow the model to learn diverse

relational patterns, proving particularly effective for capturing spatial–affective correlations in the valence dimension [43].

Finally, 1D global average pooling aggregates the sequence into a fixed-length vector, $p = 1/s \sum_{i=1}^s h_i$, where s is the sequence length and h_i denotes the i -th timestep feature. A fully connected layer then maps this representation to the predicted scores, $v = W_v p + b_v$ and $a = W_a p + b_a$, yielding the final continuous valence and arousal values.

2.3. Emotional Color Visualization Algorithm

The Emotional Color Visualization Algorithm (ECV-Algo) receives sequential V-A predictions from the GSMB modules. It generates a smooth RGB color sequence by first performing personalized mapping from V-A values to the HSV space, then preprocessing the V-A sequences, and finally applying intensity-driven dynamic interpolation in the HSV space. The resulting HSV sequence is subsequently converted to RGB for direct rendering on vehicle ambient lighting hardware.

ECV-Algo incorporates personalized mapping rules to account for individual differences in emotion–color associations. The Big Five personality traits (BFPT) include openness, conscientiousness, extraversion, agreeableness, and neuroticism. These traits form the framework for personality analysis and are determined through standardized NEO-PI-R questionnaire scores. Each dominant personality trait is assigned a distinct set of mapping rules from normalized V-A values ($V, A \in [-1, 1]$) to HSV components. Figure 3 illustrates the resulting color distributions across the five traits. Compared to traditional fixed emotion–color mapping methods, the personalized mapping rule proposed in this paper does not significantly improve objective visualization metrics. However, its main advantage lies in its ability to generate unique color sequences tailored to different personality traits, thus producing colors that better reflect an individual’s subjective emotional experience and preferences. The mapping rules primarily adjust hue based on valence, saturation and value based on arousal intensity, and have trait-specific parameters. For example, in the high openness profile, hue and auxiliary components (saturation or value) are computed as follows:

$$H = \begin{cases} 2\pi(1 - V') & V \geq 0 \\ 7\pi/6 + \pi/6(1 - V') & V < 0 \end{cases} \quad (2)$$

$$X = \begin{cases} k_1 + 0.2A' & A' \geq 0 \\ k_2 - 0.2A' & A' < 0 \end{cases} \quad (3)$$

where H represents hue and X represents saturation or value. For saturation, $k_1 = 0.8$ and $k_2 = 0.4$. For value, $k_1 = 0.7$ and $k_2 = 0.5$. All value components are constrained within the range $[0.3, 1.0]$ to ensure the visibility of the generated colors.

The raw V-A sequences from GSMB are preprocessed to reduce noise and improve mapping robustness. Outliers are first removed using a Z-score threshold of 3.0 over a 50-frame sliding window. Savitzky–Golay filtering (polynomial order 3, window length 21) is subsequently applied to attenuate minor fluctuations while preserving emotional trends. Finally, linear normalization scales the filtered values to the target HSV ranges to accommodate variations in emotional intensity across different music pieces. Dynamic frame interpolation is performed independently on H, S, and V channels to generate temporally coherent transitions. Linear interpolation is used between consecutive

frames. The interpolation density is controlled by current arousal intensity to reflect emotional dynamics:

$$N_d = \min\{N_{\max}, \max[N_{\min}, N_{\text{base}} + \beta|A(t)|]\} \quad (4)$$

where N_{base} is the base frame count and β is the dynamic interpolation coefficient. During periods of intense emotional fluctuation, higher values of $|A|$ produce denser interpolations to reflect more gradual changes. The final HSV sequence is converted to RGB using standard color space transformation for direct output to vehicle ambient lighting hardware.

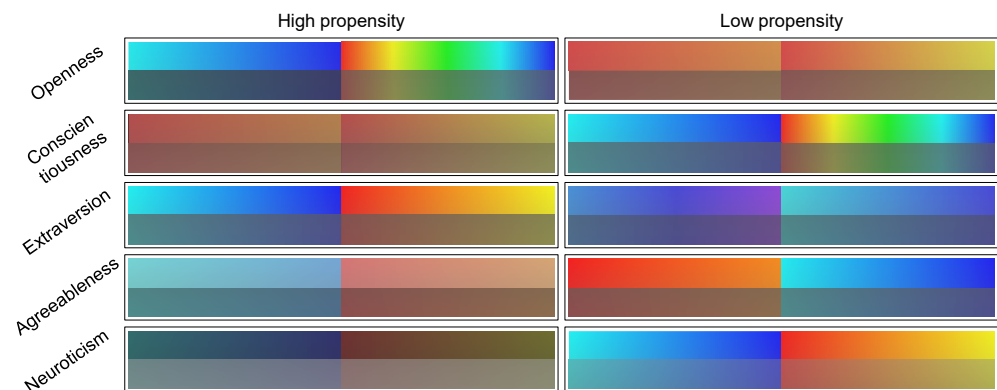


Figure 3. Personalized emotion-to-color mappings for dominant Big Five personality traits. The figure presents five color bands corresponding to the five dominant Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) from top to bottom. Hue varies primarily with valence (positive on the right half, negative on the left half), while saturation and value are modulated by arousal intensity, demonstrating personalized emotion-to-color mapping rules.

3. Experiments and Results

3.1. Experimental Setup

The proposed DC-MEPV framework is tested on the Database for Emotional Analysis in Music (DEAM) [41] and Popular Music with Emotional Annotations (PMemo) [42]. The DEAM dataset contains 1802 music excerpts, each annotated by at least seven participants under the V-A model, providing both static labels and dynamic annotations at 0.5 s intervals. The PMemo dataset includes 794 tracks with static and dynamic (0.5 s resolution) V-A annotations from at least ten annotators per song. To enhance annotation reliability, the initial 15 s of each track is excluded from dynamic labeling in both datasets. In this study, 1200 songs from DEAM and all 794 songs from PMemo are used. The selection of 1200 songs from the full DEAM dataset (out of 1802) was made to balance dataset sizes between DEAM and PMemo for equitable computational training while preserving representativeness through random sampling; this subset maintains fairness in comparison with baselines, which were evaluated under consistent conditions on the selected data. The datasets are randomly split into training and test sets in an 8:2 ratio. On an NVIDIA RTX 3090 GPU, DC-MEPV processes each 0.5 s audio frame in approximately 15 ms (excluding preprocessing), corresponding to a real-time factor of 0.03, which supports its feasibility for real-time deployment in intelligent vehicle cockpits. Ground-truth dynamic emotional trajectories are computed as the per-frame average of multiple annotators' V-A labels.

Audio preprocessing begins by segmenting each track into non-overlapping 0.5 s segments to support continuous emotion prediction. A pre-emphasis filter is applied, followed by extraction of log-Mel spectrograms using Torchaudio. The spectrograms are generated with a frame length of 50 ms, a hop length of 10 ms, and 128 Mel filter banks. DC-MEPV is implemented in PyTorch (2.4.1) and trained using the Adam optimizer with

a learning rate of 0.0001, a batch size of 64, and 100 training epochs. The hybrid loss was implemented with weights $\lambda_e = \lambda_c = 1.0$, giving equal importance to both objectives. A pre-emphasis filter with a coefficient of 0.97 was applied during audio preprocessing. To ensure reproducibility and robustness, all experiments were repeated three times with different random seeds, and the average performance metrics were reported. Detailed architectural hyperparameters are summarized in Table 1. Note that the batch size is not included in the table. To promote dynamic consistency in predicted emotional trajectories, a hybrid loss function is adopted that combines mean squared error (ρ_e) with the concordance correlation coefficient (ρ_c). Previous studies [40] have shown that MSE alone may overlook prediction–ground truth correlations, which can lead to discontinuities in regression outputs. The hybrid loss encourages global trend alignment and helps escape local optima. It is formulated as

$$\mathcal{L} = \lambda_e \cdot \rho_e + \lambda_c \cdot (1 - \rho_c), \quad (5)$$

where λ_e and λ_c are weighting hyperparameters. The term $(1 - \rho_c)$ penalizes weak correlations, as $\rho_c \in [-1, 1]$ measures agreement between predicted and ground-truth sequences.

Table 1. Model architecture hyperparameters of the proposed DC-MEPV framework, detailing the layer configurations and output dimensions for both valence and arousal channels.

Model Layer		Hyperparameters							
Input		[128,51]							
MSEFB-A		GFDM-A $\times 2$		Maxpool1d	GFDM-A	Maxpool1d	GFDM-A	GFCM-A $\times 2$	
Output Size		[256,51]	[64,1]	[256,51]	[1024,25]	[1024,12]	[2048,12]	[1024,12]	[512,12]
Filter Number/ Stride, Kernel Size		[64,1][128,3] [32,5][32,7]	[128,1][256,3] [64,5][64,7]	[2,2]	[256,1][512,3] [128,5][128,7]	[2,2]	[512,1][1024,3] [256,5][256,7]	[128,1][256,3] [64,5][64,7]	[128,1][256,3] [64,5][64,7]
MSEFB-V		GFDM-V $\times 2$		Maxpool2d	GFDM-V $\times 2$	Maxpool2d	GFDM-V	GFCM-V $\times 2$	
Output Size		[8,128,51]	[16,128,51]	[16,64,25]	[32,64,25]	[64,64,25]	[64,32,12]	[128,32,12]	[512,12]
Filter Number/ Stride, Kernel Size		[2,1][4,3] [1,5][1,1]	[4,1][8,3] [2,5][2,1]	[2,1,3]	[8,1][16,3] [4,5][4,1]	[16,1][32,3] [8,5][8,1]	[2,1,3]	[32,1][64,3] [16,5][16,1]	[32,1][64,3] [8,1][16,3] [4,5][4,1]
GSMB		Bi-LSTM		Multi-head attention			Global pool 1D		FC
Output Size		[12,256]		[12,256]			[256]		1
Filter Number/ Stride, Kernel Size		/		/			/		1

3.2. Music Emotion Recognition Evaluation

Music emotion recognition performance is evaluated using Concordance Correlation Coefficient (CCC), Root Mean Square Error (RMSE), and Pearson Correlation Coefficient (PCC). CCC evaluates the agreement between predicted and ground-truth emotional trajectories by integrating accuracy and correlation. RMSE measures absolute prediction error, while PCC quantifies linear correlation. These three metrics collectively provide a comprehensive evaluation of precision and temporal consistency in dynamic emotion recognition tasks. We demonstrate the superiority of DC-MEPV by comparing it with state-of-the-art dynamic MER methods on the DEAM and PMemo datasets, including Musi-ABC [20], BCRSN [22], LSTM-GRU [24], CBSA [43] and MERT [44]. As shown in Table 2, DC-MEPV consistently achieves the lowest RMSE and the highest PCC and CCC scores across both valence and arousal dimensions. It delivers relative CCC improvements of at least 9.9% on valence and 10.4% on arousal over the strongest baseline, highlighting its enhanced capability in modeling complex and dynamic emotional patterns.

Table 2. Performance comparison on DEAM and PMemo datasets.

Model	Valence (DEAM/PMemo)			Arousal (DEAM/PMemo)		
	RMSE↓	PCC↑	CCC↑	RMSE↓	PCC↑	CCC↑
CBSA	0.229/0.143	0.417/0.481	0.484/0.433	0.236/0.136	0.689/0.553	0.660/0.545
Musi-ABC	0.231/0.145	0.429/0.481	0.481/0.432	0.234/0.133	0.696/0.545	0.651/0.537
BCRSN	0.223/0.138	0.429/0.493	0.488/0.443	0.217/0.133	0.636/0.508	0.641/0.496
LSTM-GRU	0.332/0.255	0.383/0.411	0.403/0.389	0.239/0.257	0.588/0.433	0.473/0.402
MERT	0.228/0.142	0.452/0.475	0.492/0.478	0.209/0.131	0.682/0.562	0.671/0.551
DC-MEPV	0.219/0.135	0.567/0.519	0.541/0.516	0.197/0.129	0.739/0.584	0.741/0.579
CNN + BiLSTM	0.244/—	0.389/—	0.319/—	0.228/—	0.694/—	0.593/—
CNN + GSMB	0.239/—	0.418/—	0.392/—	0.221/—	0.705/—	0.645/—
MSFEB + BiLSTM	0.217/—	0.526/—	0.520/—	0.206/—	0.737/—	0.725/—
SC-MSFEB + GSMB	0.222/—	0.539/—	0.512/—	0.201/—	0.731/—	0.722/—

SC: Single-Channel.

Based on a standard BiLSTM model, we conducted systematic ablation experiments on the DEAM dataset to explicitly isolate the contributions of the proposed MSFEB and GSMB modules. Four configurations were compared: (1) ordinary CNN feature extraction + BiLSTM (baseline), (2) ordinary CNN + GSMB, (3) MSFEB + BiLSTM, and (4) single-channel MSFEB + GSMB. The results are shown in Table 2. Compared with the CNN+BiLSTM baseline, introducing GSMB alone improves the CCC by 22.9% on valence and 8.8% on arousal, demonstrating the benefit of global sequence modeling. Replacing ordinary CNN with MSFEB further yields substantial gains (CCC +63.0% on valence and +22.3% on arousal relative to baseline), highlighting the advantage of multi-scale feature extraction tailored to V-A dimensions. The single-channel MSFEB + GSMB variant continues to improve performance, but the full dual-channel DC-MEPV achieves the best results, with an additional +5.7% CCC on valence and +2.6% on arousal over the single-channel version. These ablation results confirm that both the MSFEB and GSMB modules contribute critically to the superior emotion perception performance of DC-MEPV, and the dual-channel design provides the final decisive improvement through dimensional decoupling.

3.3. Emotional Color Visualization Evaluation

To comprehensively evaluate the quality of music-driven ambient lighting sequences generated by ECV-Algo, we propose three novel metrics specifically aiming at human visual perception, hardware compatibility, and acousto-optic synergy, including color transition smoothness, color-emotion alignment, and color complexity. These metrics were chosen for their direct relevance to practical lighting applications: CIEDE2000 variation rate aligns with human perceptual color differences, correlation coefficients quantify emotion-visual synchronization supported by audiovisual perception studies, and unique color ratio with joint entropy ensures expressive yet non-overwhelming visualizations. We randomly selected 120 audio tracks from the PMemo dataset for our experiment to ensure that the four quadrants of the V-A emotional space are evenly covered to represent different emotional dynamics. For consistency and reproducibility, the high-openness personality mapping rule is used throughout the evaluation. Results are visualized in Figure 4.

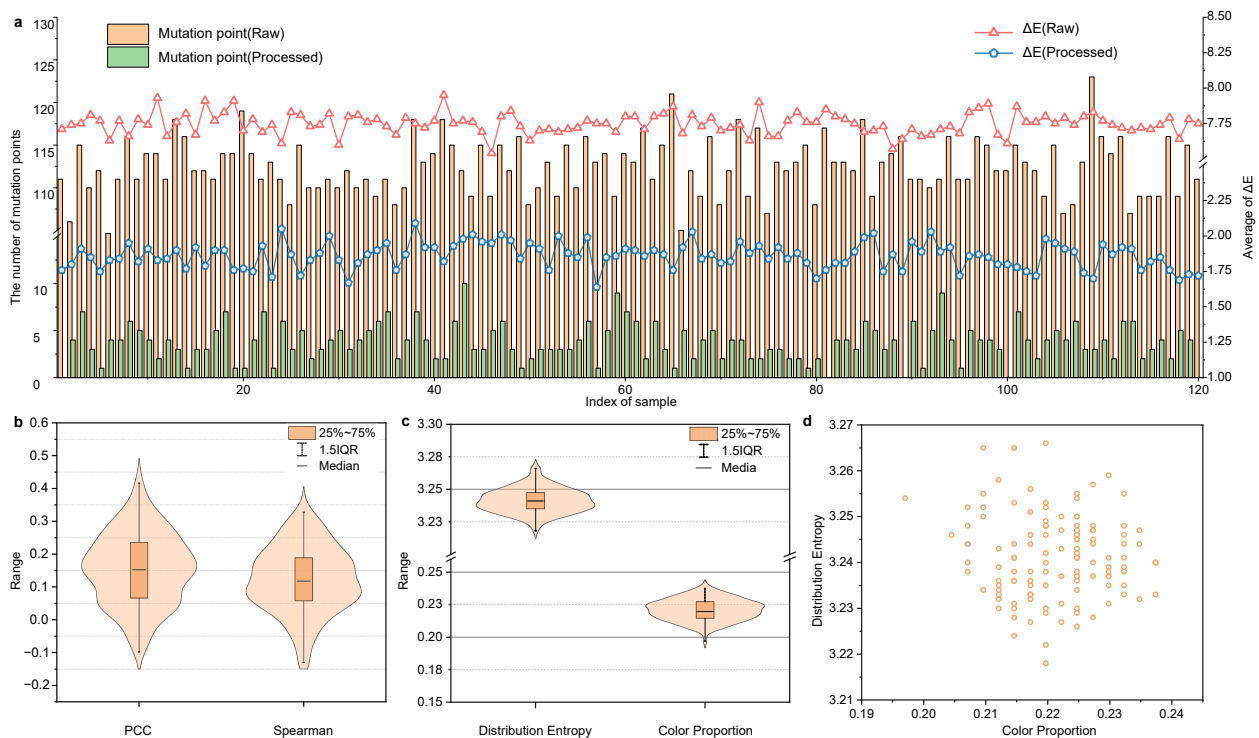


Figure 4. Experimental Results on Color Transition Smoothness, Color–Emotion Alignment, and Color Complexity. (a) Box plots of CIEDE2000 variation rate (left) and number of abrupt frames per song (right), comparing sequences with and without gradient processing. (b) Distribution of Pearson (PCC) and Spearman correlation coefficients across 120 test songs. (c) Box plots of unique color ratio and HSV joint entropy. (d) Scatter plot showing the relationship between unique color ratio (x-axis) and joint entropy (y-axis) for individual songs.

Color transition smoothness assesses the perceptual fluency of frame-to-frame changes by measuring the CIEDE2000 color difference (ΔE_{00}) and its variation rate across consecutive frames. The CIEDE2000 formula accounts for non-linear human perception in lightness, chroma, and hue:

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2} + R_T \frac{\Delta C'}{k_C S_C} \frac{\Delta H'}{k_H S_H}, \quad (6)$$

where $\Delta L'$, $\Delta C'$, and $\Delta H'$ denote differences in lightness, chroma, and hue, respectively. k_L , k_C , k_H are parametric coefficients (typically set to 1). S_L , S_C , S_H are weighting functions. R_T is a rotation term correcting hue direction interactions. As shown in Figure 4a, sequences processed by ECV- Algo exhibit an average CIEDE2000 variation rate of 1.6–2.0 and fewer than 10 abrupt frames (exceeding a perceptual threshold of 5) per song. In contrast, sequences without preprocessing and dynamic interpolation show variation rates of 7.6–7.9 and frequent noticeable jumps. These results confirm the effectiveness of outlier removal, Savitzky–Golay filtering, and intensity-driven HSV interpolation in producing perceptually smooth color transitions.

Color–emotion alignment evaluates semantic coherence between emotional shifts in music and corresponding color changes. It is quantified by PCC and Spearman rank correlation between inter-frame V–A Euclidean distances $d_{aff}(i) = \sqrt{(V_{i+1} - V_i)^2 + (A_{i+1} - A_i)^2}$ and their associated CIEDE2000 color differences $d_{col}(i)$. The distribution across the 120 test songs yields PCC values from -0.11 to 0.32 and Spearman coefficients from -0.13 to 0.40 , as shown in Figure 4b. Although absolute values are modest, they are consistent and expected under the constrained subspace of the high-openness mapping rule. The

stable trends validate that ECV-Algo reliably preserves monotonic and linear relationships between emotional dynamics and visual evolution, supporting robust audiovisual synchronization within personalized frameworks.

Color complexity measures the expressive richness of the generated sequences while avoiding visual overload. It comprises two complementary indicators: the proportion of unique colors (distinct clusters relative to total frames) and the joint entropy of the HSV distribution $H = -\sum_{b=1}^B p_b \log_2 p_b$. Results in Figure 4c show unique color ratios of 20–24% (well above the 10% over-smoothing threshold) and joint entropy values of 3.22–3.26 (within the perceptually comfortable 3–5 range). The scatter plot in Figure 4d further illustrates a dispersed yet balanced distribution between the two measures, confirming that ECV-Algo achieves diverse yet coherent color expression that effectively reflects musical emotional fluctuations. Overall, the three metrics collectively demonstrate that ECV-Algo can generate visually smooth, semantically aligned, and expressively rich ambient lighting sequences. This validates the effectiveness of personalized mapping, sequence preprocessing, and dynamic interpolation in real-time acousto-optic applications of intelligent cockpits. These quantitative results are further supported by prior user studies on personalized emotion-to-color associations, which consistently report high subjective preference scores for continuous and personality-aware visualizations in affective lighting applications [45].

3.4. Visualization of Musical Emotion

To demonstrate the practical effects of the DC-MEPV framework, we randomly selected 16 songs from the PMemo dataset. First, we extracted the first 60 s of each song, downsampling them in 0.5 s frames. Then, we generated a spectrogram for each frame using a Fast Fourier Transform, and finally a Mel spectrogram using a Mel filter bank. Figure 5a,b show the spectrogram and Mel spectrogram of a representative song, with frequency on the y-axis and frame on the x-axis, alongside an amplitude-based color bar on the right. These spectra generate V-A values through a music emotion perception network, and then visualize the emotion in RGB color sequences according to the music emotion-to-color mapping rules created by ECV-Algo. This intuitively demonstrates the ambient lighting effect in a smart cockpit. The visualization results are presented in Figure 5c. Figure 5d displays the time series color strip, where rows represent time steps (0–60 s) and columns represent song index (0–16), revealing smooth color transitions synchronized with emotional fluctuations. The mapping rules displayed in the lower right corner highlight the innovation of ECV-Algo. Each unique V-A value pair corresponds to a single color, while minimizing the color difference between adjacent emotional states, ensuring the coherence and aesthetic appeal of the visual output.

These results not only validate the generalization ability of the DC-MEPV framework in handling diverse popular music (covering different genres and emotional distributions) on the PMemo dataset, but also highlight its superiority in cross-modal emotion transfer. Compared with traditional methods, this framework achieves dynamic and continuous audiovisual synergy and contributes to a greater sense of immersion [37,46]. In smart cockpit scenarios, this personalized audiovisual synergy is expected to improve the driving experience, alleviate emotional fatigue, and provide a new paradigm for multimodal interaction in fields such as music therapy and affective computing.

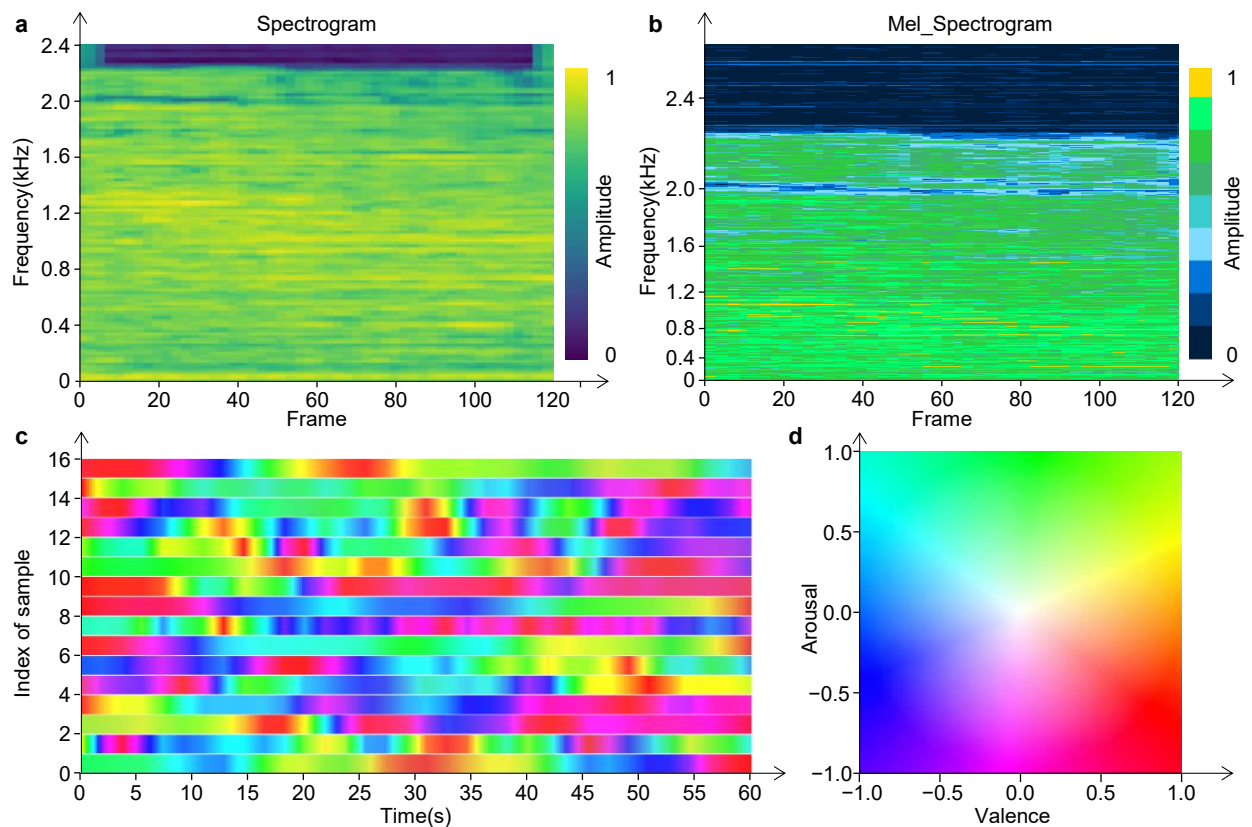


Figure 5. Visualization of music-to-emotional-color sequences.

4. Conclusions and Future Work

This paper presents DC-MEPV, a dual-channel-assisted framework for music emotion perception and visualization tailored for ambient lighting in intelligent vehicle cockpits. The framework effectively decouples valence and arousal dimensions through multi-scale feature extraction and global sequence modeling, showing superior performance on the DEAM and PMemo datasets with substantial improvements in CCC, PCC, and RMSE. The proposed ECV-Algo further realizes a personalized, continuous mapping from emotional sequences to RGB color output, resulting in smooth transitions, semantically consistent variations, and rich, expressive lighting effects, which has been validated through three new perceptual metrics and intuitive visualizations across different musical tracks. The DC-MEPV framework exhibits strong potential for broader applications in affective time-series modeling beyond intelligent vehicle cockpits. For example, its dual-channel multi-scale feature extraction and global sequence modeling may be extended to real-time driver emotion monitoring in advanced driver-assistance systems (ADAS) or wearable affective computing devices. Furthermore, the personalized ECV-Algo may be adapted for music therapy and smart home environments, where continuous emotion-to-color visualization may support therapeutic scenarios and multimodal human–machine interaction.

Despite the advances achieved with DC MEPV, limitations remain in the proposed implementation. Personalization that relies solely on the Big Five personality traits may overlook the influence of cultural, background, or real-time physiological factors on emotion–color associations. Additionally, what requires further investigation is the computational efficiency of embedded vehicle hardware and subjective evaluation in real-world driving scenarios. Future work will extend personalization by incorporating multimodal user feedback and cultural factors, optimize the framework for real-time deployment on resource-constrained systems, and conduct on-road user studies to assess impacts on driver

emotion regulation and road safety. These perspectives are expected to further promote the development of human-centered intelligent cockpit technology.

Author Contributions: W.S.: Conceptualization, Methodology, Writing—original draft. X.M.: Conceptualization, Funding, Supervision. S.L.: Writing—Original Draft, Visualization. Z.Z.: Methodology, Coding. J.L.: Writing—Review and Editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Science and Technology Project of Guangxi, grant number AA23062066.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available in the main text or the from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, W.; Zhang, B.; Wang, P.; Sun, C.; Zeng, G.; Tang, Q.; Guo, G.; Cao, D. Visual-attribute-based emotion regulation of angry driving behaviors. *IEEE Intell. Transp. Syst. Mag.* **2021**, *14*, 10–28. [\[CrossRef\]](#)
2. Liu, Y.; Xu, Q.; Guo, H.; Zhang, H. A type-2 fuzzy approach to driver-automation shared driving lane keeping control of semi-autonomous vehicles under imprecise premise variable. *Chin. J. Mech. Eng.* **2022**, *35*, 46. [\[CrossRef\]](#)
3. Braun, M.; Weber, F.; Alt, F. Affective automotive user interfaces—reviewing the state of driver affect research and emotion regulation in the car. *ACM Comput. Surv.* **2021**, *54*, 137. [\[CrossRef\]](#)
4. Lamas, D.; Loizides, F.; Nacke, L.; Petrie, H.; Winckler, M.; Zaphiris, P. *Human-Computer Interaction—INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, 2–6 September 2019; Proceedings, Part III*; Springer: Cham, Switzerland, 2019; p. 11748.
5. Li, W. Research on the Influence, Recognition, and Regulation Methods of Driver Emotional Behavior for Intelligent Vehicle Cockpits. Ph.D. Thesis, Chongqing University, Chongqing, China, 2021.
6. Jothimani, S.; Premalatha, K. MFF-SAUG: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. *Chaos Solitons Fractals* **2022**, *162*, 112512. [\[CrossRef\]](#)
7. Nalini, N.; Palanivel, S. Music emotion recognition: The combined evidence of MFCC and residual phase. *Egypt. Inform. J.* **2016**, *17*, 1–10. [\[CrossRef\]](#)
8. Panda, R.; Malheiro, R.; Paiva, R.P. Audio features for music emotion recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *14*, 68–88. [\[CrossRef\]](#)
9. Markov, K.; Matsui, T. Music genre and emotion recognition using Gaussian processes. *IEEE Access* **2014**, *2*, 688–697. [\[CrossRef\]](#)
10. Wang, J.C.; Yang, Y.H.; Wang, H.M.; Jeng, S.K. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *Proceedings of the 20th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 89–98.
11. Xianyu, H.; Li, X.; Chen, W.; Meng, F.; Tian, J.; Xu, M.; Cai, L. SVR based double-scale regression for dynamic emotion prediction in music. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2016; pp. 549–553.
12. Wu, B.; Zhong, E.; Horner, A.; Yang, Q. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the 22nd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 117–126.
13. Panda, R.; Malheiro, R.; Paiva, R.P. Novel audio features for music emotion recognition. *IEEE Trans. Affect. Comput.* **2018**, *11*, 614–626. [\[CrossRef\]](#)
14. Xu, J.; Hu, L.; Qiao, R.; Hu, Y.; Tian, Y. Music-emotion EEG coupling effects based on representational similarity. *J. Neurosci. Methods* **2023**, *398*, 109959. [\[CrossRef\]](#)
15. Mion, L.; De Poli, G. Score-independent audio features for description of music expression. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 458–466. [\[CrossRef\]](#)
16. Cai, L.; Ferguson, S.; Lu, H.; Fang, G. Feature selection approaches for optimising music emotion recognition methods. *arXiv* **2022**, arXiv:2212.13369. [\[CrossRef\]](#)

17. Zheng, W.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*; IEEE: Piscataway, NJ, USA, 2015; pp. 827–831.
18. Han, D.; Kong, Y.; Han, J.; Wang, G. A survey of music emotion recognition. *Front. Comput. Sci.* **2022**, *16*, 166335. [\[CrossRef\]](#)
19. Li, X.; Xianyu, H.; Tian, J.; Chen, W.; Meng, F.; Xu, M.; Cai, L. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2016; pp. 544–548.
20. Yang, J. Musi-ABC for predicting musical emotions. *IEEE Access* **2023**, *11*, 79455–79465. [\[CrossRef\]](#)
21. Han, X.; Chen, F.; Ban, J. Music emotion recognition based on a neural network with an inception-gru residual structure. *Electronics* **2023**, *12*, 978. [\[CrossRef\]](#)
22. Dong, Y.; Yang, X.; Zhao, X.; Li, J. Bidirectional convolutional recurrent sparse network (BCRSN): An efficient model for music emotion recognition. *IEEE Trans. Multimed.* **2019**, *21*, 3150–3163. [\[CrossRef\]](#)
23. Malheiro, R.; Panda, R.; Gomes, P.; Paiva, R.P. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Trans. Affect. Comput.* **2016**, *9*, 240–254. [\[CrossRef\]](#)
24. Angelica, F.; Hydro, R.; Sulistiya, Z.D.; Muliono, Y.; Prasetyo, S.Y. DeepLyric: Predicting Music Emotions through LSTM-GRU Hybrid Models with Regularization Techniques. *Procedia Comput. Sci.* **2023**, *227*, 973–980. [\[CrossRef\]](#)
25. Zhao, J.; Ru, G.; Yu, Y.; Wu, Y.; Li, D.; Li, W. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
26. Pridmore, R.W. Music and color: Relations in the psychophysical perspective. *Color Res. Appl.* **1992**, *17*, 57–61. [\[CrossRef\]](#)
27. Nagamachi, M. Kansei engineering: A new ergonomic consumer-oriented technology for product development. *Int. J. Ind. Ergon.* **1995**, *15*, 3–11. [\[CrossRef\]](#)
28. Watanuki, K.; Shinoda, H.; Muramatsu, K. Color emotion under LED light system controlling color rendering property. In *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*; American Society of Mechanical Engineers: New York, NY, USA, 2015; Volume 57540, p. V011T14A040.
29. Hsiao, S.W.; Hsu, C.F.; Tang, K.W. A consultation and simulation system for product color planning based on interactive genetic algorithms. *Color Res. Appl.* **2013**, *38*, 375–390. [\[CrossRef\]](#)
30. Jiang, Y.; Zhao, J. Subjective evaluation of the correlation of pop music emotion and lighting color. In *Proceedings of the 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
31. Hsiao, S.W. Fuzzy set theory on car-color design. *Color Res. Appl.* **1994**, *19*, 202–213. [\[CrossRef\]](#)
32. Palmer, S.E.; Schloss, K.B.; Xu, Z.; Prado-León, L.R. Music–color associations are mediated by emotion. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8836–8841. [\[CrossRef\]](#)
33. Nguyen, V.L.; Kim, D.; Lim, Y. Emotion-based music visualization using LED lighting control system. *J. Korea Game Soc.* **2017**, *17*, 45–52. [\[CrossRef\]](#)
34. Hsiao, S.W.; Chen, S.K.; Lee, C.H. Methodology for stage lighting control based on music emotions. *Inf. Sci.* **2017**, *412*, 14–35. [\[CrossRef\]](#)
35. Sroykham, W.; Iramina, K.; Angsuwatanakul, T.; Wongsawat, Y. Investigating the Effects of Color Light on Brain Activity Using NIRS: Implications for Emotion Regulation and Cognitive Function. In *Proceedings of the 2023 15th Biomedical Engineering International Conference (BMEiCON)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
36. Zhou, L.; Fu, T.; Miao, W.; Lin, Y. Lighting Intervention with Different Colors on Emotion: A Bayesian Network Meta-Analysis. In *Proceedings of the 2022 19th China International Forum on Solid State Lighting & 2022 8th International Forum on Wide Bandgap Semiconductors (SSLCHINA: IFWS)*; IEEE: Piscataway, NJ, USA, 2023; pp. 222–227.
37. Peng, Y.; Zhang, Z.; Jiang, Z.; Li, C.; Fan, C.; Zheng, M.; Wu, X. Driver emotion regulation via ambient lighting and music: An EEG-based ergonomic evaluation for intelligent cockpit. *Alex. Eng. J.* **2026**, *138*, 51–65. [\[CrossRef\]](#)
38. Shilandari, A.; Marvi, H.; Khosravi, H.; Wang, W. Speech emotion recognition using data augmentation method by cycle-generative adversarial networks. *Signal Image Video Process.* **2022**, *16*, 1955–1962. [\[CrossRef\]](#)
39. Ahmed, M.R.; Islam, S.; Islam, A.M.; Shatabda, S. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *Expert Syst. Appl.* **2023**, *218*, 119633. [\[CrossRef\]](#)
40. Weninger, F.; Ringeval, F.; Marchi, E.; Schuller, B. Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio. In *Proceedings of the KI 2016: Advances in Artificial Intelligence*; Friedrich, G., Helmert, M., Wotawa, F., Eds.; Lecture Notes in Artificial Intelligence; AAAI Press: Palo Alto, CA, USA, 2016; Volume 9904, pp. 310–315.
41. Aljanaki, A.; Yang, Y.H.; Soleymani, M. Developing a benchmark for emotional analysis of music. *PLoS ONE* **2017**, *12*, e0173392. [\[CrossRef\]](#)

42. Zhang, K.; Zhang, H.; Li, S.; Yang, C.; Sun, L. The PMemo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 135–142.
43. Zhong, Z.; Wang, H.; Su, G.; Liu, L.; Pei, D. Music emotion recognition based on the integration of CNN-BiLSTM and self-attention models. *J. Comput. Eng. Appl.* **2023**, *59*, 10.
44. Kang, J.; Herremans, D. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models. *arXiv* **2025**, arXiv:2502.03979. [[CrossRef](#)]
45. Jue, J.; Ha, J.H. Exploring the Relationships between Personality and Color Preference. *Front. Psychol.* **2022**, *13*, 1065372. [[CrossRef](#)] [[PubMed](#)]
46. Chen, S.; Chen, Z.; Hong, J.; Zhuang, X.; Su, C.; Ding, Z. Exploring the relationship between audio–visual perception in Fuzhou universities and college students’ attention restoration quality using machine learning. *Front. Psychol.* **2025**, *16*, 1572426. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.