# Dual-Stream Transformer with Psychoacoustic Synergy for Dynamic Music Emotion Recognition

1st Hanchen Jiang
*State Key Laboratory of Media Audio & Video*
*(Communication University of China)*
*Ministry of Education*
Beijing, China
jhc@mails.cuc.edu.cn

2nd Qin Zhang
*State Key Laboratory of Media*
*Convergence and Communication*
*Communication University of China*
Beijing, China
zhangqin@cuc.edu.cn

***Abstract*—Dynamic Music Emotion Recognition (DMER) aims to track continuous emotional variations in music, yet machine predictions still lag behind human perception, which stems from a fundamental scientific issue: conventional acoustic features such as Mel spectrograms obscure spectral details critical for psychoacoustic cues like sensory dissonance and temporal fine structure. In addition, prevalent RNN-based models struggle to capture the long-range dependencies of musical narratives. To bridge this gap, we propose the Psychoacoustic-Informed Dual-Stream Transformer (PD-Former). The method introduces Cochleogram features to simulate basilar-membrane responses, capturing physiological texture cues that complement the acoustic structure information provided by Mel spectrograms. A dual-stream convolutional architecture processes these heterogeneous features independently before synergistic fusion, and a Transformer further models long-range temporal dependencies. Experiments on the DEAM dataset show that PD-Former achieves state-of-the-art performance while remaining lightweight. Ablation studies further validate the complementarity of psychoacoustic and acoustic features, the necessity of dual-stream fusion, and the superiority of the Transformer in capturing long-range dependencies. Our model achieves notable RMSE reductions—12.5% in Valence and 15.8% in Arousal over the acoustic-only baseline, and 5.6% and 2.1% respectively over state-of-the-art benchmarks on the DEAM dataset.**

***Index Terms*—Dynamic Music Emotion Recognition, Psychoacoustics, Cochleogram, Transformer**

## I. Introduction

Music is a universal medium for emotional expression, yet machine understanding of musical emotion lags behind human perception. Music Emotion Recognition (MER) seeks to enable machines to comprehend complex musical affect, promising transformative applications in recommendation systems, music therapy, and human-computer interaction.

However, musical emotion is inherently dynamic. Unlike static music emotion recognition, which reduces complexity by assigning a single label to an entire track, dynamic music emotion recognition tracks continuous emotional evolution within the Valence-Arousal (VA) space [1]. This frame-level granularity is essential for applications requiring real-time affective feedback.

Despite progress, high-precision continuous prediction remains challenging due to the complexity of auditory cognition. We identify three critical limitations in existing approaches: (1) **Feature Representation Gap.** Most works rely on acoustic features such as Mel Spectrograms [2]. While acoustic features scale approximate pitch perception, they fundamentally lack the biophysical fidelity to model the basilar membrane's nonlinear compression and temporal fine structure [3]. Consequently, they obscure critical psychoacoustic cues such as sensory dissonance (roughness), a primary physiological driver of emotional arousal. In contrast, these cues are explicitly preserved by the Cochleogram's auditory nerve simulation. Since emotion perception relies on these psychoacoustic mechanisms [4], standard engineered features fail to fully characterize the underlying neural responses [5]. (2) **Feature Fusion Gap.** Approaches often fuse heterogeneous features via simple concatenation. This ignores differing physical scales and abstraction levels, often causing dominant features to overshadow complementary information and increasing optimization difficulty. (3) **Temporal Modeling Gap.** Musical emotion depends on long-range contextual contrasts. Traditional Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, struggle to capture these dependencies due to inherent sequential processing limitations.

To address these gaps, we propose the Psychoacoustic-Informed Dual-Stream Transformer (PD-Former), designed to emulate hierarchical auditory processing. Our contributions are: (1) **Psychoacoustic-Informed Feature Extraction.** We introduce the Cochleogram to MER. Extracted alongside Mel Spectrograms, this simulates basilar membrane responses via Gammatone filterbanks, creating a bimodal space that captures both engineered acoustic patterns and physiological cues. (2) **Dual-Stream Architecture.** We propose a Dual-Stream Convolutional Neural Network (CNN) where features are processed through independent encoders. This preserves modality-specific characteristics and promotes deeper mining of complementary information before fusion. (3) **Transformer-Based Temporal Modeling.** Replacing traditional RNNs, we employ a Transformer encoder [6]. Its self-attention mechanism efficiently captures long-range emotional dependencies across distant musical events.

Extensive evaluations on the DEAM benchmark [7] demonstrate that PD-Former achieves state-of-the-art performance. Ablation studies further validate the synergy of cochlear

features, dual-stream fusion, and Transformer modeling.

## II. Related Work

### A. Feature Representation Methods

Effective feature representation is critical for MER. While the Mel spectrogram is the *de facto* standard, used effectively by CNNs (e.g., VGG, ResNet), it is an engineered representation based on logarithmic scales that does not model the nonlinear perceptual characteristics of the Human Auditory System (HAS). The Cochleogram addresses this by simulating basilar membrane responses via Gammatone filterbanks, which match auditory nerve tuning curves and capture nonlinear frequency decomposition on the Equivalent Rectangular Bandwidth (ERB) scale. Russo et al. first showed its superiority over traditional features in music classification [8], with similar gains observed in Speech Emotion Recognition (SER).

Psychoacoustic features have recently entered DMER. Zhang et al.'s DAMFF combined MFCCs and Cochleograms using multi-scale fusion, achieving RMSE scores of 0.340 (valence) and 0.240 (arousal) on DEAM [9]. However, DAMFF employs early fusion strategy (concatenation), which ignores the features' different structures (log vs. ERB scale) and forces generic, suboptimal kernels, thus limiting the exploitation of complementary information. In contrast, computer vision domains use dual-stream networks (e.g., for RGB and optical flow) with parallel branches and high-level fusion [10]. This allows modality-specific learning but remains underexplored in MER.

### B. Temporal Dependency Modeling

Modeling temporal dependencies in DMER has been dominated by RNN-based approaches, especially LSTM and BiLSTM, often in CNN-RNN hybrids [11]. Coutinho's deep LSTM achieved a 0.372 (V) / 0.234 (A) RMSE baseline on DEAM [12], and Malik et al. used stacked CNN-BiLSTMs [13]. However, RNNs' sequential processing struggles to capture the critical long-range dependencies inherent in music.

The Transformer offers a compelling alternative, using its self-attention mechanism to create direct connections between all sequence positions, regardless of temporal distance. Its success in NLP, SER, and physiological signal analysis has led to its adoption in MER [14]. Zhang et al. used a dual-scale attention Transformer [15], while Chen et al.'s Transformer encoder achieved a 0.247 (V) / 0.224 (A) RMSE on DEAM [16], substantially improving on RNN baselines.

Despite this progress, current Transformer-based DMER models suffer from two limitations: (1) they rely almost exclusively on single Mel spectrogram inputs, neglecting valuable psychoacoustic features, and (2) effective fusion of heterogeneous feature streams within the Transformer architecture remains underexplored [17].

## III. PD-Former

We propose the Psychoacoustic-Informed Dual-Stream Transformer (PD-Former), an end-to-end deep learning architecture designed to learn dynamic music emotion from parallel acoustic and psychoacoustic-informed features. This section details the model's architecture and constituent components.

### A. Overall Architecture

The overall architecture of PD-Former (see Fig. 1) is an end-to-end dual-stream temporal network consisting of five stages:

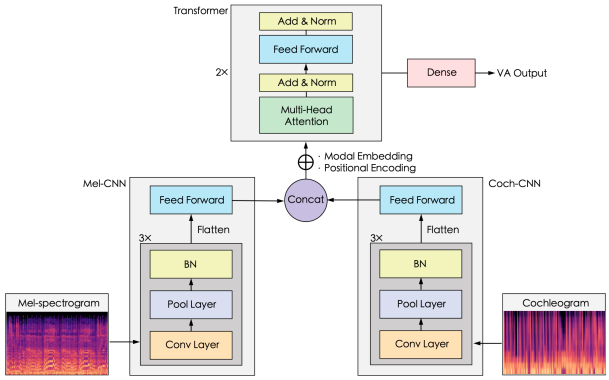


Fig. 1. The overall architecture of the proposed PD-Former.

1) **Parallel Feature Extraction.** We extract two parallel time-frequency representations from the raw audio: the standard Mel spectrogram and the psychoacoustic-informed cochleogram.
2) **Dual-Stream CNN Encoding.** We segment both feature maps into fixed-length (0.5s) clips and process each clip through a TimeDistributed wrapper before feeding it into its respective CNN encoder. The Mel-CNN and Coch-CNN streams operate independently without weight sharing, which preserves each modality's unique characteristics while learning high-dimensional feature representations.
3) **Temporal Feature Fusion.** After CNN encoding, we concatenate the feature vectors from both streams at each time step to form a unified temporal feature sequence.
4) **Transformer-based Temporal Modeling.** We augment the fused sequence with positional encoding before passing it to a Transformer encoder stack, whose self-attention mechanism captures long-range temporal dependencies.
5) **VA Regression Output.** A fully connected regression head processes the Transformer output to predict frame-by-frame Valence-Arousal (VA) values. We optimize the model by minimizing 1-CCC loss, which maximizes the Concordance Correlation Coefficient between predictions and ground truth.

### B. Psychoacoustic-Informed Feature Extraction

As shown in Fig. 2, PD-Former employs dual-stream inputs to capture both standard acoustic patterns and psychoacoustic-informed non-linear responses. We now detail the extraction and preprocessing procedures for each feature type.

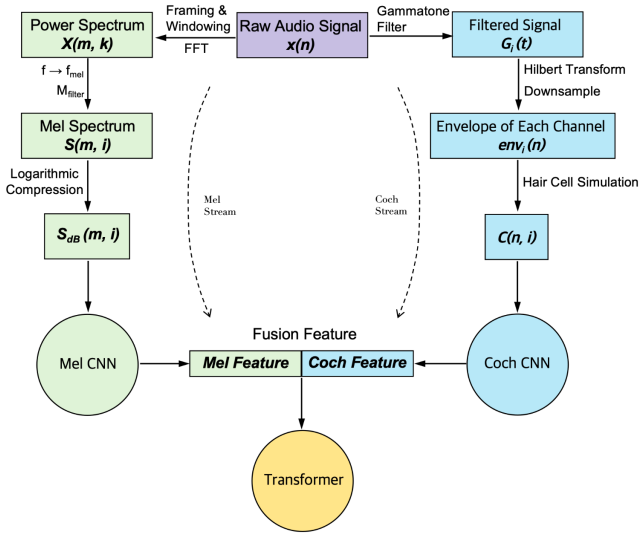


Fig. 2. Overview of the PD-Former's dual-stream feature encoding.

*1) Mel Spectrogram:* The Mel spectrogram provides a time-frequency representation that models the human ear's non-linear frequency perception [18].

We first frame and window the input audio signal $x(n)$ to compute the power spectrum:

$$X(m,k) = \left| \sum_{n=0}^{N_{\text{fft}}-1} x(n+mH)w(n)e^{-j2\pi kn/N_{\text{fft}}} \right|^2, \quad (1)$$

where $m$ is the frame index, $w(n)$ is the Hann window, $N_{\text{fft}} = 2048$ is the FFT size, $H = \lfloor f_s/88 \rfloor$ is the hop length, and $f_s = 44.1$ kHz is the sampling rate. This hop length yields a temporal resolution of 88 Hz.

Next, we convert from the linear frequency scale to the perceptually-motivated Mel scale:

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \quad (2)$$

We then apply a triangular filter bank $M_{\text{filter}}(i,k)$ with 96 filters to map the power spectrum onto Mel-frequency bands:

$$S(m,i) = \sum_{k=0}^{N_{\text{fft}}/2} M_{\text{filter}}(i,k)X(m,k), \quad (3)$$

where $i \in \{1,\ldots,96\}$ indexes the Mel bands.

Finally, to model the ear's logarithmic intensity response, we convert to decibels:

$$S_{\text{dB}}(m,i) = 10\log_{10}\left(S(m,i) + \epsilon\right), \quad (4)$$

where $\epsilon = 10^{-8}$ ensures numerical stability.

*2) Cochleogram:* The cochleogram employs an auditory periphery model based on a Gammatone filterbank to simulate the basilar membrane's frequency decomposition mechanism [19].

The cochleogram stream is designed to serve as a lightweight, psychoacoustic-informed complement [20] to the dense acoustic representation of the Mel spectrogram. In contrast to the 96 bands used for the Mel spectrogram, we design 11 Gammatone filters with center frequencies distributed on the Equivalent Rectangular Bandwidth (ERB) scale, spanning 20 Hz to 22,050 Hz. This 11-band psychoacoustic representation is consistent with prior successful work, such as Du et al. [21], which demonstrated the efficacy of modulation-filtered cochleograms for emotion recognition. The time-domain impulse response of the $i$-th filter is:

$$g_i(t) = t^3 e^{-2\pi b_i t}\cos(2\pi f_i t + \phi_i)u(t), \quad (5)$$

where $f_i$ is the center frequency, $b_i = 1.019 \times \text{ERB}(f_i)$ is the bandwidth parameter, $\text{ERB}(f) = 24.7(4.37f/1000+1)$ is the Equivalent Rectangular Bandwidth, and $u(t)$ is the unit step function.

We filter the input signal $x(t)$ through the Gammatone filterbank and extract the envelope of each channel using the Hilbert transform:

$$\begin{aligned} y_i(t) &= x(t) * g_i(t), \\ \text{env}_i(t) &= \sqrt{y_i^2(t) + \mathcal{H}\{y_i(t)\}^2}, \end{aligned} \quad (6)$$

where $*$ denotes convolution and $\mathcal{H}$ represents the Hilbert transform.

To maintain temporal consistency with the Mel spectrogram, we downsample the envelope signals to 88 Hz:

$$\text{env}_i^{\text{down}}(n) = \text{Resample}\left(\text{env}_i(t), \frac{88}{f_s}\right), \quad (7)$$

where $f_s = 44.1$ kHz is the sampling rate.

To model the compressive non-linearity of inner hair cells, we apply logarithmic scaling:

$$C(n,i) = 20\log_{10}\left(\max(\text{env}_i^{\text{down}}(n), \epsilon)\right), \quad (8)$$

where $\epsilon = 10^{-8}$ ensures numerical stability.

*3) Feature Segmentation:* The extraction pipelines produce two time-frequency feature maps: a $(96 \times 2640)$ Mel spectrogram and an $(11 \times 2640)$ cochleogram. Following the DEAM dataset protocol, we extract a 30-second segment (15–45 s) from each 60-second audio clip, corresponding to 2640 frames at 88 Hz. We then segment each feature map into a sequence of 60 time steps, where each step contains a 2D patch spanning 0.5 seconds (44 frames). This yields input shapes of $(96, 44, 1)$ for Mel spectrograms and $(11, 44, 1)$ for cochleograms.

This segmentation produces a 5D tensor of shape $(\text{Batch}, 60, F_{\text{bins}}, 44, 1)$, where $F_{\text{bins}} \in \{96, 11\}$ denotes the number of frequency channels. This design enables us to apply 2D CNNs independently to each spatial patch via a `TimeDistributed` wrapper, extracting local spectrotemporal features, before the Transformer models long-range temporal dependencies across the 60-step sequence.

### C. Dual-Stream CNN Feature Encoder

To effectively process two heterogeneous feature representations—standard acoustic features (Mel spectrogram) and psychoacoustic-informed features (Cochleogram)—we design

a Dual-Stream Convolutional Neural Network (CNN) encoder. Rather than directly concatenating features at the input stage—which would force a single CNN to learn two fundamentally different data distributions simultaneously—we employ parallel processing streams.

Our dual-stream design enables two parallel CNN branches ($\text{CNN}_{\text{mel}}$ and $\text{CNN}_{\text{coch}}$) to independently extract modality-specific representations. $\text{CNN}_{\text{mel}}$ extracts standard time-frequency energy patterns from the Mel spectrogram, while $\text{CNN}_{\text{coch}}$ decodes non-linear auditory information embedded in the Cochleogram that simulates human perceptual characteristics. This specialized processing more effectively captures complementary information from both modalities.

To preserve temporal information during spatial feature extraction, we wrap both CNN encoders with a `TimeDistributed` layer. The `TimeDistributed` layer applies the same CNN operation independently to each time step in the sequence. Each block comprises a `Conv2D` layer, a `BatchNormalization` layer for training stability, and a `MaxPooling2D` layer for spatial dimensionality reduction. Subsequently, `GlobalAveragePooling2D` compresses each 2D feature map into a 1D vector at every time step. Finally, a feed-forward layer projects each vector into a compact 32-dimensional embedding space. The outputs from both CNN encoders are then concatenated to form a unified feature sequence. This sequence serves as input to the Transformer module for temporal modeling and affective information fusion.

### *D. Transformer-based Temporal Fusion*

The Dual-Stream CNN encoder produces a fused feature sequence. Each feature vector at time step $t$ represents only the instantaneous state, lacking global musical context. Music emotion is inherently time-dependent: the current emotional state results from the interplay between past melodic evolution and anticipated future events (e.g., an impending climax).

Unlike RNNs or LSTMs, the Transformer's self-attention mechanism computes dependencies between arbitrary time steps in parallel, regardless of temporal distance. This property is particularly advantageous for modeling complex musical structures.

The feature sequence is then processed by a stack of two Transformer encoder blocks. Each encoder block comprises two core sub-layers:

1) **Multi-Head Self-Attention (MHSA):** This mechanism forms the core of the Transformer architecture. The MHSA layer enables each position in the sequence to simultaneously attend to all other positions. This is achieved using eight parallel attention heads. Each head independently computes query ($Q$), key ($K$), and value ($V$) matrices to generate weighted context vectors. This enables learning different types of temporal dependencies across distinct representation subspaces (e.g., rhythmic patterns in one head, melodic contours in another).
2) **Position-wise Feed-Forward Network (FFN):** Following the attention sub-layer, each time step's output independently passes through an FFN. This FFN comprises two linear transformations with a ReLU activation. This provides non-linear processing capability to further refine features.

After each sub-layer (MHSA and FFN), we apply residual connections and layer normalization. These components are crucial for training deep Transformer models and mitigating vanishing gradient problems.

After passing through two encoder layers, we obtain the final context-aware sequence $F_{\text{out}} \in \mathbb{R}^{T \times D_{\text{model}}}$. Finally, a dense layer is applied independently to each time step, projecting from 64 dimensions to the two-dimensional valence–arousal (V–A) emotion space with linear activation:

$$Y_{\text{pred}} = \text{Linear}(F_{\text{out}}) \in \mathbb{R}^{T \times 2}, \tag{9}$$

where $Y_{\text{pred}}$ represents the predicted dynamic V–A emotion sequence for the entire music segment.

### *E. Optimization Objective*

The ultimate objective of our model is to predict a dynamic valence-arousal (V-A) emotion sequence that exhibits high concordance with human-annotated ground truth. For such continuous time-series regression tasks, standard L1 (MAE) or L2 (MSE) loss functions have inherent drawbacks. They solely penalize the magnitude of point-wise errors, completely disregarding the dynamic trends and correlation within the sequence. A model optimized solely on MSE may produce an overly "flat" prediction curve; while the average error might be low, it fails to capture the critical emotional fluctuations inherent in the music.

To address this limitation, we adopt the *Concordance Correlation Coefficient* (CCC) [22] as our primary optimization objective. The CCC is the gold standard for assessing the agreement between two sequences (ground truth $Y$ and prediction $\hat{Y}$), as it simultaneously penalizes deviations in correlation, mean, and variance.

For a ground truth sequence $Y$ and a predicted sequence $\hat{Y}$, the CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_Y\sigma_{\hat{Y}}}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\mu_Y - \mu_{\hat{Y}})^2}, \tag{10}$$

where $\mu_Y$ and $\mu_{\hat{Y}}$ are the respective means of $Y$ and $\hat{Y}$; $\sigma_Y^2$ and $\sigma_{\hat{Y}}^2$ are their variances; and $\rho$ is the Pearson correlation coefficient between them.

The CCC ranges from $-1$ to 1, where $+1$ signifies perfect concordance. Our objective is to maximize the CCC. Consequently, we define our loss function $\mathcal{L}_{\text{CCC}}$ as:

$$\mathcal{L}_{\text{CCC}} = 1 - \text{CCC}. \tag{11}$$

Our model is required to predict both valence and arousal dimensions simultaneously. Thus, we compute $CCC_V$ for valence and $CCC_A$ for arousal independently and define the final loss as their average:

$$\mathcal{L}_{\text{Total}} = 1 - \frac{1}{2}(CCC_V + CCC_A). \tag{12}$$

By minimizing this loss function, we compel the model to learn not only the absolute values of the valence and arousal dimensions (governed by the $\mu$ and $\sigma$ terms), but also their correct temporal dynamics (governed by the $\rho$ term).

## IV. Experiments

To systematically evaluate PD-Former, we conduct comprehensive experiments on the DEAM benchmark to address three key questions: (1) Does our model achieve state-of-the-art performance? (2) Can it effectively capture dynamic emotional trajectories? (3) Are the proposed components—psychoacoustic features, dual-stream architecture, and Transformer modeling—necessary and synergistic? We present dataset details and implementation, experimental results and visualization, comparison with state-of-the-art methods, and ablation studies.

### A. Dataset and Metrics

*1) Dataset:* All our experiments were conducted on the DEAM (Dynamic Emotion in Music) dataset. DEAM is one of the most authoritative and widely used public datasets in the DMER field. It contains 1,802 music clips of diverse styles. Each track is provided with continuous *Valence* and *Arousal* annotations (every 0.5 seconds), with values ranging from $-1$ to 1. These dynamic annotations were provided by multiple annotators using the JOPS (joystick) tool and were averaged to reflect the collective group perception.

Following standard practice in previous DMER work, we segmented the features and corresponding annotations of each song into fixed-length sequences of 60 time steps. If the remaining part of a song was shorter than 60 time steps, padding was applied. We split the 1,802 tracks into training (80%), validation (10%), and test (10%) sets at the song level, ensuring a fair comparison with SOTA methods.

*2) Evaluation Metrics:* Besides CCC, following the standard practice in DMER, we selected the Root Mean Square Error (RMSE) as the primary metric for evaluation and comparison. RMSE measures the magnitude of the differences between predicted values ($\hat{y}_i$) and ground truth values ($y_i$). It is a standard metric for evaluating regression accuracy and is particularly sensitive to larger errors. A lower RMSE indicates better model performance. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{13}$$

where $N$ is the total number of samples, $y_i$ is the ground truth value, and $\hat{y}_i$ is the predicted value.

### B. Implementation Details

We implemented the model using TensorFlow 2.5.0. Features were extracted using `librosa` (for Mel spectrograms) [23] and `pycochleagram` (for Cochleograms), consistent with the parameters detailed in Section III. Each feature map was normalized per-sample using Min-Max normalization before being fed into the model. We used the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$ and applied gradient clipping (clipnorm $= 1.0$) for training stability. The model was trained with a batch size of 32, optimizing the $1 - \text{CCC}$ loss function. All Dropout layers in the architecture were set to a rate of 0.2 to mitigate overfitting.

### C. Training Dynamics and Stability

Figure 3 illustrates the convergence process of the PD-Former model over 70 epochs. The validation loss curve closely tracks the training loss curve throughout the process, indicating that the model generalizes well and does not suffer from significant overfitting. Based on the minimum validation loss, we selected the model from Epoch 56 as the best-performing model for final evaluation. When evaluated on the independent test set, this final model achieved a CCC of 0.663 for Valence and 0.715 for Arousal.

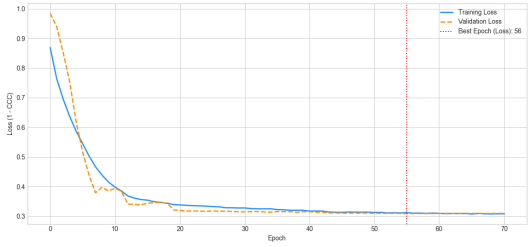


Fig. 3. Model convergence plot showing training loss and validation loss $(1 - \text{CCC})$ over 70 epochs. The best model was selected at Epoch 56.

To explore the model's robustness boundaries and validate the chosen parameters, we conducted a stability analysis on key hyperparameters, with results shown in Table I.

TABLE I
Hyperparameter Stability Analysis (RMSE).

| Parameter | Value | Valence (RMSE) ↓ | Arousal (RMSE) ↓ |
|---|---|---|---|
| *Learning Rate* | $1 \times 10^{-3}$ | 0.241 (Unstable) | 0.198 (Unstable) |
| | $1 \times 10^{-4}$ | **0.168** | **0.139** |
| | $1 \times 10^{-5}$ | 0.195 (Slow) | 0.169 (Slow) |
| *Transformer Layers* | 1 | 0.179 | 0.150 |
| | **2** | **0.168** | **0.139** |
| | 3 | 0.173 | 0.145 |
| | 6 | 0.177 (Overfit) | 0.149 (Overfit) |

The results demonstrate that PD-Former is robust and not overly sensitive to hyperparameter fluctuations. The model maintains consistent performance across moderate settings (e.g., comparing $N_T = 2$ and 3), confirming its stability. Performance only degrades at distinct logical boundaries: optimization instability occurs at excessive learning rates ($10^{-3}$), and overfitting emerges only when model depth is significantly increased ($N_T = 6$).

### D. Qualitative Analysis

To intuitively evaluate PD-Former's capability to capture emotional dynamics, we selected a representative instrumental music sample (DEAM track ID = 131) for visual analysis. As shown in Figure 4, our model's predictions effectively fit the

overall trends of the ground truth annotations for both Valence and Arousal dimensions. For instance, around Time Step 6, as the music transitions to a new theme with increased percussion and loudness, the model accurately predicts the sharp increase in VA values. Similarly, after Time Step 36, as the music enters a calm coda, the model's predictions correctly track the gradual decrease.

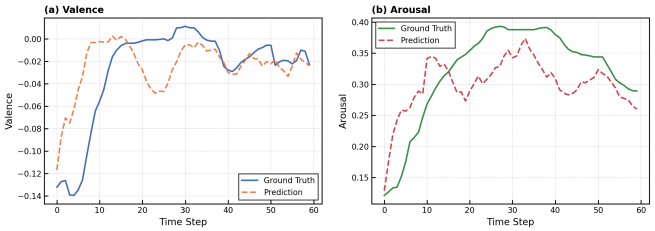


Fig. 4. Visualization of dynamic emotion prediction for a test sample from the DEAM dataset (track ID = 131). The plots compare the model's (a) Valence and (b) Arousal predictions (dashed line) against the ground truth (solid line) over 60 time steps.

Discrepancy analysis reveals key model characteristics. First, at Time Step 6, a new musical theme was introduced in the original audio. The model's rapid response at Time Step 8 is faster than the annotation, which likely reflecting the inherent perceptual-motor latency of human annotation rather than a model error. Second, the transient fluctuation at Time Step 20 is a genuine model error, indicating over-sensitivity to minor acoustic features (at time step 20, the original audio undergoes a chord change), whereas human hearing possesses stronger emotional inertia. Therefore, we assess the model's predictions as highly credible for tracking emotional trends, but its transient errors suggest future correction methods, such as introducing temporal smoothing or explicitly modeling emotional inertia.

### *E. Robustness to Temporal Scales*

While DMER typically focuses on fine-grained tracking (e.g., 0.5s in DEAM), practical applications often require diverse temporal resolutions. For instance, music therapy playlist generation relies on longer-term emotional stability. To evaluate PD-Former's robustness across different temporal scales, we conducted experiments with varying observation windows: 5s, 10s, and 15s. The results are shown in table II.

TABLE II
PERFORMANCE COMPARISON ACROSS DIFFERENT TEMPORAL WINDOWS (RMSE). AGGREGATED VIA MEAN POOLING.

| Window Size | Valence (RMSE) ↓ | Arousal (RMSE) ↓ |
|---|---|---|
| 0.5s (Original) | 0.168 | 0.139 |
| 5.0s | 0.142 | 0.112 |
| 10.0s | 0.125 | 0.098 |
| 15.0s | 0.118 | 0.091 |

Since the DEAM dataset is annotated at 0.5s intervals, we generated ground truth labels for longer windows by computing the arithmetic mean of the corresponding 0.5s annotations. This averaging process aligns with the concept of emotional inertia, filtering out transient annotation noise to reflect the dominant emotional state of a musical phrase. Similarly, we aggregated the model's frame-level predictions using the same non-overlapping window averaging strategy before calculating the RMSE.

### *F. Comparison with SOTA Methods*

To comprehensively and objectively evaluate the effectiveness of our proposed **PD-Former**, this section presents a quantitative comparison with both baseline and state-of-the-art (SOTA) models on the DEAM dataset. The compared models include both classical RNN-based architectures and advanced attention-based frameworks, as follows:

- **Deep LSTM-RNN (Baseline):** A foundational DMER model [12] employing stacked LSTMs to capture emotional temporal dependencies in music.
- **DAMFF (SOTA-1):** A Dual Attention-based Multi-scale Feature Fusion model [9] designed to aggregate emotional features across multiple scales.
- **Transformer Encoder (SOTA-2):** A Transformer-based architecture [16] that replaces traditional RNNs for sequence modeling in MER.
- **BCRSN (SOTA-3):** A Bidirectional Convolutional Recurrent Sparse Network [24], representing one of the strongest published results on the DEAM dataset.

*1) Performance Comparison and Analysis:* Table III reports the RMSE performance of all models on the DEAM dataset.

TABLE III
RMSE PERFORMANCE COMPARISON ON THE DEAM DATASET. LOWER IS BETTER.

| Model | Valence (RMSE) ↓ | Arousal (RMSE) ↓ |
|---|---|---|
| Deep LSTM-RNN (Baseline) | 0.372 | 0.234 |
| DAMFF (SOTA-1) | 0.340 | 0.240 |
| Transformer Encoder (SOTA-2) | 0.247 | 0.224 |
| BCRSN (SOTA-3) | 0.178 | 0.142 |
| **PD-Former (Ours)** | **0.168** | **0.139** |

Our PD-Former achieves the lowest RMSE on both Valence and Arousal, clearly outperforming the Baseline and SOTA-1. This demonstrates the superior feature extraction and temporal modeling capability of our framework. Although both PD-Former and SOTA-2 adopt Transformer architectures, our model significantly surpasses SOTA-2, primarily due to the proposed dual-stream feature input. Unlike SOTA-2, which relies solely on acoustic features, PD-Former introduces the Cochleogram, simulating the human auditory system's nonlinear perception. This psychoacoustic feature complements the Mel spectrogram and provides richer, more discriminative representations for the Transformer. SOTA-3 is a highly optimized RNN-based model that achieves remarkable efficiency through sparse encoding. However, our self-attention–based PD-Former achieves comparable or better accuracy without any recurrent structure, showing that Transformer-based modeling—when paired with psychoacoustic front-ends—can effectively capture long-range emotional dependencies in music.

*2) Model Complexity and Efficiency Analysis:* To further evaluate computational efficiency, we analyze PD-Former's parameter count and architectural balance. PD-Former contains only 95,234 trainable parameters. The CNN-based feature extraction modules (Mel-CNN and Coch-CNN) contribute 52,928 parameters (55.6%), while the Transformer temporal module accounts for 42,306 parameters (44.4%). Our 95K parameter model is highly efficient. Compared to SOTA, it avoids the heavy BPTT costs of LSTMs (Baseline) and the sparse regularization of BCRSN (SOTA-3). Furthermore, its CNN front-end (53K params) effectively prepares features for the lightweight Transformer module (42K params), enhancing efficiency over standard Transformer encoders (SOTA-2) and redundant multi-scale architectures (SOTA-1).

In summary, PD-Former achieves SOTA-level accuracy while maintaining exceptional computational efficiency, validating the effectiveness of our proposed Psychoacoustic-Informed Dual-Stream architecture for efficient and high-performance DMER.

### G. Ablation Study

We designed a series of rigorous ablation studies to thoroughly investigate the necessity and effectiveness of each key component within our proposed PD-Former.

*1) Experimental Setup:* We designed the following four variant models to compare against our full model:

- **Replacing the Temporal Module (PD-BiLSTM):** This variant replaces the complete Transformer temporal fusion module with a Bidirectional Long Short-Term Memory (BiLSTM) network. This is intended to validate the superiority of the Transformer in capturing long-range dependencies.
- **Mel-Spectrogram Only (Mel-Former):** This variant removes the Cochleogram feature stream, using only the Mel spectrogram as input, which is then processed by a single-stream CNN encoder and the Transformer.
- **Cochleogram Only (Coch-Former):** Contrary to Mel-Former, this variant removes the Mel spectrogram stream, using only the Cochleogram feature stream. These two variants together are used to validate the complementarity and necessity of the dual-stream features.
- **Early Fusion (PD-EarlyFusion):** This variant removes the two parallel CNN encoders. The Mel spectrogram and Cochleogram are simply concatenated at the input stage and then fed directly into the Transformer. This simulates an "early fusion" strategy and is used to demonstrate the effectiveness of our designed CNN encoders.

*2) Results and Analysis:* Table IV presents the RMSE results of the ablation study.

TABLE IV
ABLATION STUDY RMSE RESULTS ON THE DEAM DATASET (LOWER IS BETTER).

| Model Configuration | Valence (RMSE) ↓ | Arousal (RMSE) ↓ |
|---|---|---|
| PD-BiLSTM | 0.185 | 0.153 |
| Mel-Former | 0.192 | 0.165 |
| Coch-Former | 0.201 | 0.171 |
| PD-EarlyFusion | 0.253 | 0.202 |
| **PD-Former** | **0.168** | **0.139** |

Compared to the Full Model, using only the Mel spectrogram (Mel-Former) or only the Cochleogram (Coch-Former) both lead to a significant drop in performance. Although the Mel spectrogram is a standard feature, it is insufficient to capture all emotional cues. The Cochleogram, while having psychoacoustic advantages, also requires the information from the Mel spectrogram as a supplement. This strongly proves that the two features are highly complementary and that our dual-stream design is rational and necessary, successfully fusing the strengths of both representations.

The performance of PD-BiLSTM is inferior to the Full Model, which uses the Transformer. This confirms one of our core hypotheses: when processing a sequence like music, which possesses complex structures and long-range dependencies, the self-attention-based Transformer is superior to BiLSTM in temporal modeling capabilities.

PD-EarlyFusion exhibits the worst performance of all variants (V: 0.253, A: 0.202). This model removes the CNN encoders, directly feeding the raw (or near-raw) feature maps into the Transformer after fusion. This results in a catastrophic performance decline. This indicates that our designed dual-stream CNN encoders are indispensable as feature extractors. They successfully abstract the features from both modalities into high-level, compact, and informative representations before the temporal fusion stage.

In summary, the ablation study systematically validates every component of our model's design. The results demonstrate that the psychoacoustic-informed Cochleogram, the dual-stream CNN feature encoders, and the Transformer temporal fusion module work synergistically, and none can be dispensed with.

## V. CONCLUSION

This paper addresses a critical limitation in Dynamic Music Emotion Recognition: the disconnect between acoustic feature extraction and human psychoacoustic perception. We propose PD-Former, a psychoacoustic-informed dual-stream architecture that achieves state-of-the-art performance on the DEAM benchmark (RMSE: 0.168 for Valence, 0.139 for Arousal) while maintaining exceptional efficiency with only 95K parameters.

Our core contributions are threefold. First, we introduce the Cochleogram as a psychoacoustic complement to standard Mel spectrograms, simulating basilar membrane responses to capture perceptual cues obscured by purely engineered features. Second, we design a dual-stream CNN architecture that processes heterogeneous feature modalities independently before fusion, enabling effective exploitation of complementary information. Third, we employ Transformer encoders to capture long-range temporal dependencies critical for modeling musical emotion dynamics, surpassing traditional RNN-based approaches.

Rigorous ablation studies validate our hypothesis: emotional perception emerges from the synergy of acoustic and psychoacoustic cues, and dual-stream processing is essential for their effective integration. Our visualization analyses demonstrate that PD-Former successfully tracks complex emotional trajectories in music, though occasional over-sensitivity to transient acoustic changes suggests opportunities for incorporating temporal smoothing or explicit emotional inertia modeling in future work.

## REFERENCES


[1] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 3, no. 3, pp. 1–30, 2012.

[3] Y. E. Kim et al., "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Utrecht, Netherlands, Aug. 2010, pp. 937–952.

[4] E. Coutinho and A. Cangelosi, "Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements," *Emotion*, vol. 11, no. 4, pp. 921–931, 2011.

[5] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, UK: Oxford University Press, 1990.

[6] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[7] A. Aljanaki, Y. H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, p. e0173392, 2017.

[8] M. Russo, L. Kraljević, M. Stella, and M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Inf. Process. Manag.*, vol. 57, no. 5, Art. no. 102270, 2020.

[9] L. Zhang, X. Yang, Y. Zhang, and J. Luo, "Dual attention-based multi-scale feature fusion approach for dynamic music emotion recognition," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Milan, Italy, Nov. 2023, pp. 207–214.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[11] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN based music emotion recognition from raw audio signal," in *Proc. 29th Int. Conf. Radioelektronika*, Pardubice, Czech Republic, Apr. 2019, pp. 1–4.

[12] E. Coutinho, G. Trigeorgis, S. Zafeiriou, and B. Schuller, "Automatically estimating emotion in music with deep long-short term memory recurrent neural networks," in *CEUR Workshop Proc.*, vol. 1436, Jan. 2015, pp. 1–3.

[13] M. Malik *et al.*, "Stacked convolutional and recurrent neural networks for music emotion recognition," arXiv preprint arXiv:1706.02292, 2017.

[14] R. Liyanarachchi, A. Joshi, and E. Meijering, "A survey on multimodal music emotion recognition," arXiv preprint arXiv:2504.18799, 2025.

[15] D. Zhang, W. You, Z. Liu, L. Sun, and P. Chen, "Personalized dynamic music emotion recognition with dual-scale attention-based meta-learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 2, pp. 1629–1637, Apr. 2025.

[16] Y. Chen, Z. Ma, M. Wang, and M. Liu, "Advancing music emotion recognition: A Transformer encoder-based approach," in *Proc. 6th ACM Int. Conf. Multimedia Asia*, Auckland, New Zealand, Dec. 2024, pp. 1–5.

[17] J. Kang and D. Herremans, "Are we there yet? A brief survey of music emotion prediction datasets, models and outstanding challenges," arXiv preprint arXiv:2406.08809, 2024.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[19] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, Dec. 1987.

[20] Z. Peng, W. He, Y. Li, Y. Du, and J. Dang, "Multi-level attention-based categorical emotion recognition using modulation-filtered cochleogram," *Appl. Sci.*, vol. 13, no. 11, Art. no. 6749, 2023.

[21] P. Du, X. Li, and Y. Gao, "Dynamic music emotion recognition based on CNN-BiLSTM," in *Proc. 2020 IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Chongqing, China, June 2020, pp. 1372–1376.

[22] B. T. Atmaja and M. Akagi, "Evaluation of error—and correlation—based loss functions for multitask learning dimensional speech emotion recognition," *J. Phys.: Conf. Ser.*, vol. 1896, no. 1, Art. no. 012004, Apr. 2021.

[23] B. McFee et al., "Librosa: Audio and music signal analysis in Python," in *Proc. SciPy*, 2015, pp. 18–24.

[24] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional convolutional recurrent sparse network (BCRSN): An efficient model for music emotion recognition," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3150–3163, 2019.