Symbolic and Acoustic: Multi-domain Music Emotion Modeling for Instrumental Music

Kexin Zhu, Xulong Zhang, Jianzong Wang$^{(\boxtimes)}$, Ning Cheng, and Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China
jzwang@188.com

Abstract. Music Emotion Recognition involves the automatic identification of emotional elements within music tracks, and it has garnered significant attention due to its broad applicability in the field of Music Information Retrieval. It can also be used as the upstream task of many other human-related tasks such as emotional music generation and music recommendation. Due to existing psychology research, music emotion is determined by multiple factors such as the Timbre, Velocity, and Structure of the music. Incorporating multiple factors in MER helps achieve more interpretable and finer-grained methods. However, most prior works were uni-domain and showed weak consistency between arousal modeling performance and valence modeling performance. Based on this background, we designed a multi-domain emotion modeling method for instrumental music that combines symbolic analysis and acoustic analysis. At the same time, because of the rarity of music data and the difficulty of labeling, our multi-domain approach can make full use of limited data. Our approach was implemented and assessed using the publicly available piano dataset EMOPIA, resulting in a notable improvement over our baseline model with a 2.4% increase in overall accuracy, establishing its state-of-the-art performance.

Keywords: Piano emotion recognition · Music information retrieval · Multi-domain analysis

1 Introduction

The emotional aspect of music, commonly known as its affective content, holds significant importance and is often regarded as the essence of musical expression. The recognition of emotions in music, known as Music Emotion Recognition (MER), has emerged as a prominent topic and crucial objective within the field of Music Information Retrieval (MIR). This recognition process assumes paramount significance due to its widespread application in various scenarios involving emotion-driven music retrieval and recommendation. Restricted by the

complexity of emotion, research on MER has encountered great difficulties [9,31]. Emotion is a very complex psychological state, and different people have different emotional thresholds [32]. This makes emotional annotation more difficult and emotional data more scarce.

The recognition and understanding of the intricate interplay between various factors within music and their impact on music emotion constitute a central concern in ongoing research on MER. Investigating this matter not only facilitates the advancement of more efficient and nuanced MER techniques but also contributes to the development of comprehensive insights into the complex nature of music emotion. Existing research usually applies disentanglement or multi-domain analysis to modeling music emotion from multiple aspects. Berardinis et al. [1] applies Music Source Separation during pre-processing and analyze the emotional content in vocal, bass, drums, and other parts separately, their proposed method shows promising performance. Zhao et al. [37] provide a new perspective by modeling music emotion with both music content and music context, their proposed method applies multi-modal analysis on audio content, lyrics, track name, and artist name of the music.

To further explore the essence of music emotion, research was also carried out on instrumental music. In the field of psychology and affective computing, Laukka et al. [18] proposed a convincing music emotion perception model for instrumental music and concluded six factors that affect music emotion: Dynamics, Rhythm, Timbre, Register, Tonality, and Structure. Those factors reflect both the acoustic characteristics and the structural characteristics of the music. Laukka's model indicates the importance of incorporating both acoustic analysis and symbolic analysis for MER. Acoustic factors such as Dynamics and Timbre are highly related to the Arousal expression of the music but are not included in the symbolic representations of music. Therefore symbolic-only methods show relatively weaker performance on Arousal detection. Structural factors such as Tonality and Structure are highly related to the Valence expression. Although those factors are included in the acoustic domain, existing acoustic analysis methods can hardly learn the structural information without extra supervision. To incorporate all the important factors, both acoustic analysis and symbolic analysis are needed.

However, most existing MER methods for instrumental music are uni-domain and fail to model music emotion from multiple aspects. Existing researches mainly apply deep-learning-based methods on the acoustic domain or uses sequence-modeling methods on the symbolic domain representations of the music. In their recent publication on emotion recognition in symbolic music, Qiu et al. [30] introduced a pioneering approach utilizing the MIDIBERT model [4], a large-scale pre-trained music understanding model. At present, no existing research on Music Emotion Recognition (MER) for instrumental music integrates both acoustic and symbolic analyses. As a result, we present an innovative method in this study that encompasses music emotion modeling from both acoustic and symbolic perspectives. Given the representative nature of piano music within the instrumental domain, we implemented and conducted an evaluation

of our proposed approach using the publicly available piano emotion dataset EMOPIA [16].

Our contribution can be summarized as follows:

– Inspired by existing psychology and affective computing research, we proposed a multi-domain emotion modeling method for instrumental music, which only needs audio input. Our method used a pre-trained transcription model to obtain symbolic representation, therefore can be used on each instrument that can be automatically transcribed.
– We designed a refined acoustic model with mixed acoustic features input and a transformer-based symbolic model. Both models showed promising performance.
– We implemented and evaluated our proposed method on the public piano emotion dataset EMOPIA [16]. Our method achieved state-of-the-art performance on EMOPIA with better consistency between Valence detection and Arousal detection performance.

2   Related Works

There have been many studies in the research field of MER. According to the different domains of focus, these studies include MER with acoustic-only and MER with symbolic-only studies. These works have promoted progress in MER, and there are also some points that can be improved.

2.1   MER with Acoustic-Only

In order to explore which part of the vocal or accompaniment music carries more emotional information, Xu et al. [36] used the sound source separation technology, combined with the 84-dimensional manual low-level features (such as Mel frequency cepstrum coefficient (MFCC), spectral center, spectral attenuation point, spectral flux, and other similar measures.), and then used a classifier to recognize music emotion. Coutinho et al. [6] extracted 65 Low-level Descriptors (LLDs) in a time window of 1 s and calculated their first-order difference to obtain a total of 130 low-level features, then calculated the mean and standard deviation of each LLD in one second, and finally formed a 260-dimensional feature vector, and then used Long Short-term Memory (LSTM) network to carry out regression prediction of dynamic V/A (Valence/Arousal) value. Fukayama et al. [8] proposed a method to adapt to aggregation by considering new acoustic signal input based on multi-stage regression. At the same time, a method of adjusting the aggregation weight is introduced to deal with the emotion caused by the new input that cannot be known in advance, and the deviation observed in the training data is utilized by using Gaussian process regression. Li et al. [19] introduced a novel approach to tackle dynamic emotion regression by leveraging Deep Bi-directional Long Short-term Memory (DBiLSTM) in a multi-scale regression framework. Moreover, the author also examined the influence of dissimilar sequence lengths between the training and prediction stages on the overall

performance of DBiLSTM. By investigating this aspect, the study aimed to gain insights into the effects of such variations on the efficacy of the model. [23] uses the CNN network that can process local information with fewer parameters and the RNN network that can process context information, that is, the CRNN structure, which uses the least parameters than Media Eval 2015.

Other methods have achieved the best results in the dynamic regression prediction of emotion at that time. Huang et al. [14] introduced the attention mechanism into the music emotion classification task, and introduced the attention layer with short-term and short-term memory units into the deep convolution neural network for music emotion classification. Different weights are allocated on different time blocks (chunks), and the song-level emotion classification prediction is obtained through fusion. Liu et al. [22] regards music emotion recognition as a multi-label classification task, and uses convolutional neural networks and spectrum diagram to complete end-to-end classification. Chen et al. [2] considered the complementarity between CNN with different structures and between CNN and LSTM, and combined multi-channel CNN with different structures and LSTM into a unified structure (Multi-channel Convolutional LSTM, MCCLSTM) to extract advanced music descriptors. Choi et al. [3] employed a pre-trained convolutional neural network (CNN) feature, which was initially trained for music auto-tagging purposes. They then successfully transferred this CNN to various music-related classification and regression tasks, showcasing its adaptability and versatility. Similarly, Panda et al. [27] introduced a collection of innovative affective audio features to enhance emotional classification in audio music. The authors observed that conventional feature extractors primarily focus on low-level timbre-related aspects, neglecting essential elements like musical form, texture, and expressive skills. To address this limitation, the authors devised a novel set of algorithms specifically designed to capture information related to music texture and expression, effectively compensating for the significant gaps in music emotion recognition research.

2.2   MER with Symbolic-Only

Previous research employed manual extraction of statistical musical characteristics, which were subsequently inputted into machine learning classifiers to forecast the emotional aspects of notated music. Grekow et al. [10] conducted an analysis on classical music in MIDI format and extracted 63 distinct features. In a similar vein, Lin et al. [20] conducted a comparative investigation involving multiple features (audio, lyrics, and MIDI) extracted from the same music. Remarkably, they discovered that MIDI features exhibited superior performance in emotion recognition. Building upon this finding, the researchers utilized the JSymbolic library [25] to extract 112 advanced music features from MIDI files. Subsequently, Support Vector Machine (SVM) was employed to classify the data. Similarly, Panda et al. [28] employed various tools to extract features from MIDI files and utilized SVM for classification purposes.

More recent studies demonstrate a growing adoption of a symbolic music encoding technique similar to MIDI [26], which is gaining popularity among

researchers. Additionally, deep learning models have emerged as the prominent approach in this field. Ferreira [7] devised a method to encode MIDI files into MIDI-like sequences, leveraging LSTM and GPT2 for sentiment classification purposes. This approach offers simplicity and efficiency. Drawing inspiration from the remarkable achievements of BERT, Chou et al. [5] introduced MidiBERTPiano, a large-scale pre-trained model utilizing CP representation. The proposed model showcases promising outcomes in various domains, including symbolic music emotion recognition. Highlighting the paramount importance of emotional expression in music's intrinsic structure, Liu et al. [30] proposed a straightforward multi-task framework for the symbolic MER task. Notably, this approach benefits from readily available labels for auxiliary tasks, eliminating the need for manual annotation of labels beyond emotion classification.

3   Methodology

The complete diagram illustrating the overall architecture of our proposed approach can be observed in Fig. 1. The structure contains two branches: the acoustic domain branch (marked in yellow) applies acoustic analysis on mixed acoustic features with a Conv-based acoustic encoder, and the symbolic domain branch (marked in blue) applies symbolic analysis on music score sequence by using a Transformer-based symbolic encoder. It is worth noting that the outputs of the two branches come from the same modality, that is, from the acoustic input, so they belong to different domains of the same modality.
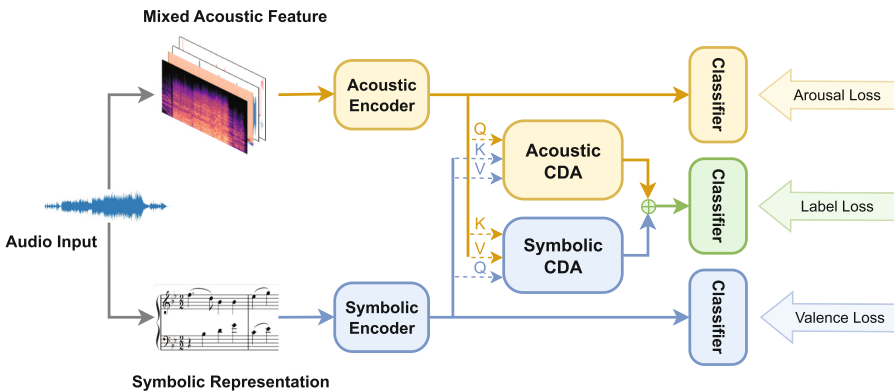

Fig. 1. The overall structure. The feature representations of the two domains are generated by the acoustic domain branch and the symbolic domain branch in the model, and the fusion process is completed in CDA.

3.1   Acoustic Domain Analysis for Arousal Modeling

For the acoustic domain analysis, we want to explicitly extract the information that relates to music emotion expressions, such as Timbre and Dynamics [18]. We use a mixed feature as input, which consists of the Mel-frequency Cepstral Coefficient (MFCC), Mel-spectrogram, Spectral Centroid (SC), and Root Mean Square Energy (RMSE) of the audio input. SC and RMSE reflect the energy distribution and changes of the audio, which is strongly correlated to music emotion expression. We use mel-spectrogram instead of STFT spectrogram because it better fits the human auditory perception process. We also calculate a 20-dimensional MFCC with librosa [24]. After these features are obtained, we resize and align them in the time dimension. The mixed feature can be obtained by splicing these features.

The processing flow of the acoustic domain branch is shown at the top of Fig. 1. We use a 2D-ConvNet module as the acoustic encoder for its great ability to encode temporal and frequency domain information simultaneously. After the feature extraction process, the extracted features are flattened and combined in the channel dimension to form the acoustic domain output. A comprehensive summary of the settings used in the experiment can be found in Table 1.

Acoustic domain analysis shows better performance on Arousal detection than symbolic domain analysis. Arousal is mainly decided by acoustic attributes such as Dynamics, Energy, and Timbre, which are not included in symbolic domain representation. Therefore we calculate an extra arousal classification loss function using Binary Cross Entropy (BCE) on the acoustic domain analysis branch during the training process.

3.2   Symbolic Domain Analysis for Valence Modeling

As mentioned above, our proposed method is designed to perform both acoustic and symbolic domain analysis with only audio input. That is to say, our symbolic part uses the automatic piano transcription module to form the symbolic domain representation instead of directly using the MIDI files in the EMOPIA dataset. This provides a common paradigm for other transcribable musical instruments. Therefore for the symbolic domain analysis branch, we use a pre-trained automatic transcription model to perform piano transcription. Specifically, we use the refined version of Onsets and Frames [11,12] proposed by Zhao et al. [38], which shows better generalizability and costs fewer computation resources. The transcribed piano score is converted into MIDI format, which includes the onset, offset, duration, and velocity of each note.

The music score is the "language" of the music and is a semantic sequence similar to natural language. Therefore the symbolic representation of the music score is similar to that of the natural language.

In this work, we use a refined MIDI-like representation for note embedding, which is shown in Fig. 2. Unlike the original MIDI-like [26] representation, we add an attribute named "harmonic" which explicitly denotes the number of sounding notes at the onset of a note. Since harmonic is an important part of

musical performance, we decide to add extra information about it. Therefore, the symbolic domain representation for a single note consists of the onset time, harmonic, velocity, time shift, and offset time of the note.
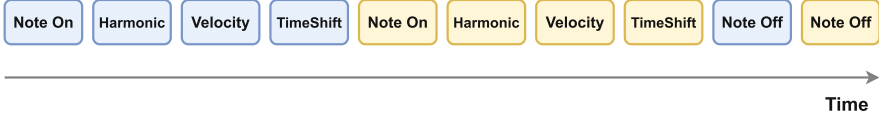

Fig. 2. The refined MIDI-like symbolic representation we used.

The structure of the symbolic domain analysis branch is shown at the bottom of Fig. 1. After the note embeddings are obtained, we input them into a Transformer encoder module [34] to extract the emotional representation of the piano score. The Transformer encoder module consists of four original Transformer encoder layers adopted in [34]. We pre-trained the encoder with the MIDI data from the MAESTRO dataset, for there are not enough samples in EMOPIA to train our Transformer encoder module.

Symbolic domain analysis mainly focuses on the high-level semantics of the note sequences, which leads to better Valence detection accuracy than acoustic analysis. As we want to make use of its advantage, we calculate an extra valence classification loss on the symbolic domain analysis branch during the training process.

3.3 Combining Symbolic and Acoustic Analysis

The final purpose of our method is to perform 4-Quadrant (4Q) classification concerning both Arousal and Valence, therefore the cross-domain feature fusion method is important. When combining extracted acoustic domain features and symbolic domain features, the Cross-domain Attention (CDA) module is used for cross-domain feature fusion. CDA has a similar mechanism to multi-head cross-modal attention [33]. In CDA module, Query and Key-Value pairs come from two different domains instead of different modalities in cross-modal attention. Each attention head can be calculated separately:

$$
\begin{aligned}
Attention(F_Q, F_K, F_V) &= softmax(\frac{F_Q(F_K)^T}{\sqrt{d}})F_V \\
&= softmax(\frac{F_\alpha W_Q(F_\beta W_K)^T}{\sqrt{d}})F_\beta W_V
\end{aligned}
\tag{1}
$$

Let F_Q, F_K, and F_V denote the vectors for *Query*, *Key*, and *Value*, respectively. Within the attention mechanism, these input vectors are obtained by multiplying the extracted features of the α and β domains, represented as F_α and F_β, with their respective learnable weight matrices W_Q, W_K, and W_V. Here,

d represents the dimension size of the Key vector. The multi-head attention can be defined as the concatenation of each individual head:

$$MultiHead(F_\alpha, F_\beta) = Concat(head_1, ..., head_H)W_O \qquad (2)$$

$$head_i = Attention(F_\alpha W_Q^i, F_\beta W_K^i, F_\beta W_V^i) \qquad (3)$$

The learnable weight matrix W_O and the number of attention heads H play crucial roles in this multi-head attention mechanism. By leveraging multiple attention heads, this mechanism effectively highlights the significant aspects of each domain, which cannot be achieved through simple concatenation alone.

As shown in Fig. 1, in each processing procedure, our model calculates the CDA mechanism twice. We calculate an acoustic cross-domain attention mechanism and a symbolic cross-domain attention mechanism separately. This bidirectional CDA fusion strategy brings higher fusing efficiency. The output of acoustic CDA and symbolic CDA are concatenated and input into a classifier for 4Q emotion classification. During the training process, we calculate a 4Q Label loss on this classifier using Cross Entropy (CE) loss function.

4 Experiments

To assess the effectiveness of our proposed model, we conducted two primary types of experiments in this study: comparative studies and ablation studies. These experiments were designed to thoroughly evaluate and analyze the performance of our model from different perspectives.

4.1 Expriments Setup

We use the EMOPIA [16] dataset, which is an open-source dataset for piano-based emotion recognition. EMOPIA contains 1087 piano clips from 387 songs, all piano clips are annotated with their MIDI files and emotion labels. As only music metadata is available, we collect all music files by their corresponding YouTube ID with the 'youtube-dl' package. Following the configuration employed in [16], the dataset was divided into train-validation-test splits with a ratio of 7:2:1, ensuring appropriate proportions for training, validation, and testing stages. However, due to the unavailability of several music pieces on YouTube, we're only able to use approximately 90% data of the whole dataset. Similarly, we not only perform the classification of 4 quadrants but also carry out the binary classification tasks of high/low Valence and high/low Arousal. For the pre-training phase of the Automatic Piano Transcription model, we utilized the MAESTRO dataset ("MIDI and Audio Edited for Synchronous TRacks and Organization") [12], encompassing a comprehensive collection of more than 200 h of meticulously paired audio and MIDI recordings.

Table 1. Acoustic Encoder Settings.

Layer	Channel	Kernel Size	Stride	Maxpooling
conv1	64	3×3	1	2×2
conv2	128	3×3	1	2×2
conv3	256	3×3	1	2×2

During the training process, the training data is divided into mini-batches with a batch size of 64. The Adam optimizer [17] is employed, utilizing a learning rate of 0.0001. To implement all experiments, the PyTorch framework [29] is utilized.

It is important to note that MIDI files from the EMOPIA dataset were not utilized in our experiments. As our proposed model exclusively takes audio files as input, our aim is to evaluate the overall performance of the complete model, including the refined AMT module.

Table 2. Comparison with symbolic-domain methods on EMOPIA.

Method	4Q	A	V
LSTM-Attn [21]+MIDI-like [26]	.684	.882	.833
LSTM-Attn [21]+REMI [15]	.615	**.890**	.746
symbolic-LR [16]	.581	.849	.651
MIDIBERT [4]	.634	/	/
MT-MIDIBERT [4,30]	.676	/	/
proposed model	**.708**	.874	**.869**

4.2 Comparative Studies

We compared our proposed model with other existing methods on the same EMOPIA dataset. To the best of our knowledge, there is no existing multi-domain piano emotion recognition research. So we compared our model with several uni-domain symbolic-domain models proposed in [16] and [30], including two models based on BLSTM and self-attention mechanism (LSTM-Attn for short) using MIDI-like and REMI symbolic representation, a linear regression model based on hand-crafted features, and two pre-trained Bert-like models. For a fair comparison, we directly used the original results announced in their works. In [4,30], valence metrics and arousal metrics are not provided, therefore are not shown in the table.

Table 2 shows the comparison between our method and the other five symbolic-domain methods. All the methods show high and similar performance on Arousal detection, which indicates that Arousal detection is a relatively simple task. Due to the strong sequence-modeling ability of our transformer-based symbolic domain model, our method shows the highest Valence detection performance and outperforms the LSTM-Attn+MIDI-like model by 3.6%. On 4Q classification metrics, our model also achieves state-of-the-art performance and outperforms the LSTM-Attn+MIDI-like model by 2.4%.

We also compared our model with two existing acoustic-domain models, one uses linear regression on hand-crafted features and the other uses a ResNet-like network. Table 3 shows the comparison between our method and the other two acoustic-domain methods. All acoustic-domain methods show strong performance on Arousal detection as well. This is in line with common sense, because Arousal is greatly affected by energy, velocity, and dynamics, and this information is evident in acoustic information. Though our method is slightly weaker on Arousal detection, it still outperforms the Short-chunk ResNet model by 3.1% on the 4Q metrics.

Table 3. Comparison with acoustic-domain methods on EMOPIA.

Method	4Q	A	V
Audio-LR [16]	.523	**.919**	.558
Short-chunk ResNet [13,35]	.677	.887	.704
proposed model	**.708**	.874	**.869**

4.3   Ablation Studies

We designed and carried out a series of ablation studies to test the effect of our improvements. In the symbolic-only model and acoustic-only model, we use our symbolic branch and acoustic branch individually in order to test the effect of combining them. In the STFT-input model, we use an STFT spectrogram as input instead of the mixed acoustic feature. In the Single-loss model, we do not calculate the extra loss on the two branches and only calculate the Label loss.

The experimental results of the ablation studies are shown in Table 4. Compared to the two uni-domain models, our cross-domain fusion strategy costs performance loss on Arousal and Valence detection. However, our proposed model outperforms these two models by over 5% on the overall 4Q accuracy metrics. This indicates that our model is able to make better decisions by considering both symbolic and acoustic information.

The STFT-input model shows huge performance loss on Arousal metrics, which proves that using mixed acoustic features can improve Arousal detection

performance. When using the STFT spectrogram as input, a deeper network is needed to extract the acoustic features. By using hand-crafted features, our method shows strong acoustic modeling ability with only three Conv layers. The single-loss model also shows over 2.5% performance loss on both 4Q and Arousal metrics, which indicates that our strategy of calculating the extra loss function works.

Table 4. Ablation studies trained and evaluated on the EMOPIA dataset.

Method	4Q	A	V
Symbolic-only	.651	.843	**.891**
Acoustic-only	.630	**.902**	.697
STFT-input	.689	.804	.871
Single-loss	.683	.845	.883
proposed model	**.708**	.874	.869

5   Conclusion

In this study, we introduce a novel multi-domain approach for piano emotion recognition, which can also be extended to other instruments with automatic transcription capabilities. Our proposed model leverages a pre-trained transcription model, enabling multi-domain analysis solely based on audio input. To the best of our knowledge, there is a lack of research specifically addressing piano emotion recognition. Our proposed model capitalizes on the complementary and redundant aspects between the acoustic and symbolic domains, leading to improved consistency in valence detection and arousal detection. Experimental results demonstrate that our proposed model surpasses the baseline approaches in terms of Valence classification and 4Q classification metrics. Moving forward, our future work will focus on designing enhanced symbolic representations for music, investigating superior cross-domain fusion strategies to enhance overall performance, and developing a universal framework for addressing the emotional aspects of transcribed musical instruments.

References

1. de Berardinis, J., Cangelosi, A., Coutinho, E.: The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability. In: Proceedings of the 21st International Society for Music Information Retrieval Conference, pp. 310–317 (2020)
2. Chen, N., Wang, S.: High-level music descriptor extraction algorithm based on combination of multi-channel cnns and lstm. In: Proceedings of the 18th International Society for Music Information Retrieval Conference, pp. 509–514 (2017)

3. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Transfer learning for music classification and regression tasks. In: 18th International Society for Music Information Retrieval Conference, pp. 141–149. International Society for Music Information Retrieval (2017)

4. Chou, Y.H., Chen, I., Chang, C.J., Ching, J., Yang, Y.H., et al.: Midibert-piano: large-scale pre-training for symbolic music understanding. arXiv:2107.05223 (2021)

5. Chou, Y.H., Chen, I., Chang, C.J., Ching, J., Yang, Y.H., et al.: Midibert-piano: Large-scale pre-training for symbolic music understanding. arXiv:2107.05223 (2021)

6. Coutinho, E., Trigeorgis, G., Zafeiriou, S., Schuller, B.: Automatically estimating emotion in music with deep long-short term memory recurrent neural networks. In: Working Notes Proceedings of the MediaEval 2015 Workshop, vol. 1436, pp. 1–3 (2015)

7. Ferreira, L., Whitehead, J.: Learning to generate music with sentiment. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, pp. 384–390 (2019)

8. Fukayama, S., Goto, M.: Music emotion recognition with adaptive aggregation of gaussian process regressors. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 71–75. IEEE (2016)

9. Gómez-Cañón, J.S., et al.: Music emotion recognition: toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Process. Mag. **38**(6), 106–114 (2021)

10. Grekow, J., Raś, Z.W.: Detecting emotions in classical music from midi files. In: Foundations of Intelligent Systems: 18th International Symposium, pp. 261–270. Springer (2009)

11. Hawthorne, C., et al.: Onsets and frames: Dual-objective piano transcription. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, pp. 50–57 (2018)

12. Hawthorne, C., et al.: Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: 7th International Conference on Learning Representations (2019)

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

14. Huang, Y.S., Chou, S.Y., Yang, Y.H.: Music thumbnailing via neural attention modeling of music emotion. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 347–350. IEEE (2017)

15. Huang, Y.S., Yang, Y.H.: Pop music transformer: beat-based modeling and generation of expressive pop piano compositions. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1180–1188 (2020)

16. Hung, H., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.: EMOPIA: a multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, pp. 318–325 (2021)

17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (2015)

18. Laukka, P., Eerola, T., Thingujam, N.S., Yamasaki, T., Beller, G.: Universal and culture-specific factors in the recognition and performance of musical affect expressions. Emotion **13**(3), 434 (2013)

19. Li, X., Tian, J., Xu, M., Ning, Y., Cai, L.: Dblstm-based multi-scale fusion for dynamic emotion prediction in music. In: 2016 IEEE International Conference on Multimedia and Expo, pp. 1–6. IEEE (2016)
20. Lin, Y., Chen, X., Yang, D.: Exploration of music emotion recognition based on midi. In: Proceedings of the 14th International Society for Music Information Retrieval Conference, pp. 221–226 (2013)
21. Lin, Z., et al.: A structured self-attentive sentence embedding. In: 5th International Conference on Learning Representations (2017)
22. Liu, X., Chen, Q., Wu, X., Liu, Y., Liu, Y.: Cnn based music emotion classification. arXiv:1704.05665 (2017)
23. Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., Jarina, R.: Stacked convolutional and recurrent neural networks for music emotion recognition. arXiv:1706.02292 (2017)
24. McFee, B., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25. Citeseer (2015)
25. McKay, C., Fujinaga, I.: jsymbolic: a feature extractor for midi files. In: Proceedings of the 2006 International Computer Music Conference (2006)
26. Oore, S., Simon, I., Dieleman, S., Eck, D., Simonyan, K.: This time with feeling: learning expressive musical performance. Neural Comput. Appl. **32**(4), 955–967 (2020)
27. Panda, R., Malheiro, R., Paiva, R.P.: Musical texture and expressivity features for music emotion recognition. In: 19th International Society for Music Information Retrieval Conference, pp. 383–391 (2018)
28. Panda, R., Malheiro, R., Rocha, B., Oliveira, A., Paiva, R.P.: Multi-modal music emotion recognition: a new dataset, methodology and comparative analysis. In: International Symposium on Computer Music Multidisciplinary Research (2013)
29. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (2019)
30. Qiu, J., Chen, C., Zhang, T.: A novel multi-task learning method for symbolic music emotion recognition. arXiv:2201.05782 (2022)
31. Ru, G., Zhang, X., Wang, J., Cheng, N., Xiao, J.: Improving music genre classification from multi-modal properties of music and genre correlations perspective. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10097241
32. Tang, H., Zhang, X., Wang, J., Cheng, N., Xiao, J.: Emomix: emotion mixing via diffusion models for emotional speech synthesis. In: 24th Annual Conference of the International Speech Communication Association (2023)
33. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, vol. 2019, p. 6558. NIH Public Access (2019)
34. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
35. Won, M., Ferraro, A., Bogdanov, D., Serra, X.: Evaluation of cnn-based automatic music tagging models. arXiv:2006.00751 (2020)
36. Xu, J., Li, X., Hao, Y., Yang, G.: Source separation improves music emotion recognition. In: Proceedings of International Conference on Multimedia Retrieval, pp. 423–426 (2014)

37. Zhao, J., Ru, G., Yu, Y., Wu, Y., Li, D., Li, W.: Multimodal music emotion recognition with hierarchical cross-modal attention network. In: 2022 IEEE International Conference on Multimedia and Expo, pp. 1–6. IEEE (2022)
38. Zhao, J., Wu, Y., Wen, L., Ma, L., Ruan, L., Wang, W., Li, W.: Improving automatic piano transcription by refined feature fusion and weighted loss. In: Proceedings of the 9th Conference on Sound and Music Technology. pp. 43–53. Springer, Cham (2023). doi: https://doi.org/10.1007/978-981-19-4703-2_4