2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features

Jiahao Zhao and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {jzhao, yoshii}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—This paper describes automatic music emotion recognition (MER) that aims to estimate the valence and arousal (V/A) scores of a piece of piano music. The emotion is multifaceted in nature; it is rendered by various features that are often mutually dependent and inherent in music composition and performance. A basic approach to MER is to train a deep neural network (DNN) that extracts latent features representing the emotion as a whole and estimates the V/A scores, using only a limited amount of audio data with imbalanced V/A annotations. Such a black-box approach, however, suffers from limited performance and interpretability. To overcome these limitations, in this paper we propose a multimodal multifaceted MER method that fuses various kinds of musically-meaningful symbolic and acoustic features extracted from both MIDI and audio data, respectively, based on the expert knowledge of musical psychology. More specifically, our method separately extracts the affective features representing the rhythm, dynamics, melody, harmony, and tone color of a piano piece as the main factors affecting the emotion and integrates them with a self-attention mechanism that can learn the complicated cross-modal relationships. The experiments using the common EMOPIA dataset showed that the proposed model achieved the state-of-the-art V/A classification accuracy of 69.2% and that the multimodal and multifaceted feature fusions contributed to the performance improvement.**

I. Introduction

Music emotion refers to the affective information conveyed through music and is considered the essence of music itself. It is thus necessary to analyze music data from both objective and subjective viewpoints. As well as automatic music transcription (AMT), music emotion recognition (MER) has been a long-standing fundamental task in the field of music information retrieval (MIR) [1], because it plays a crucial role in emotion-based personalized music retrieval and recommendation systems. The abstract nature of the emotion and the expensive cost of collecting subjective annotations, however, pose a challenge to the development of MER methods.

Music emotion is multifaceted and multimodal in nature; various features of music composition and performance may implicitly interact to render an emotion in a sophisticated manner unique to the musical piece. This makes deep learning (DL) methods [2]–[4] hard to learn a consistent mapping from a specific type of music representation (e.g., score, lyrics, or audio) to the overall emotion of the piece, resulting in the limited performance of the valence and arousal (V/A) estimation. To mitigate this problem, for instance, Berardinis et al. [5] separately extracted affective acoustic features from the vocal, bass, drum, and the other parts of a music signal obtained with a source separation method. Zhao et al. [6] separately extracted text and acoustic features from the text (i.e., lyrics and track and artist names) and audio data and fused them one by one with a cross-modal attention mechanism. However, these existing methods still have some problems: on the one hand, these methods have high requirements for experimental data and can only be applied to specific music; on the other hand, the performance and interpretability of these methods are still unsatisfactory.

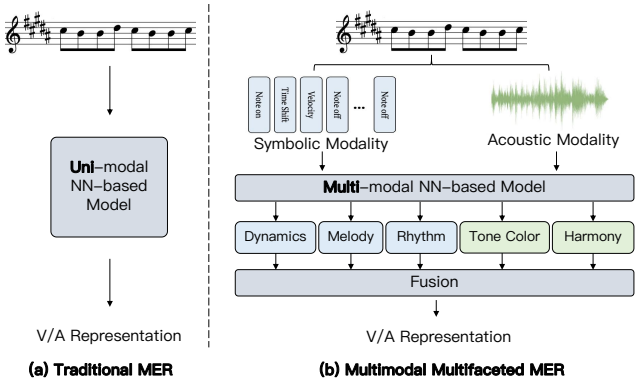

Fig. 1. Comparison between the conventional and proposed approaches.

For interpretable MER, we refer to the investigations of affective features in the fields of music psychology and affective computing. Laukka et al. [7], for example, proposed a human perception model for instrumental music and found that six factors, i.e., dynamics, rhythm, timbre, register, tonality, and structure, characterize the emotional expression of music. Renato [8] revisited this issue from a computational perspective and found eight factors useful for MER, i.e., melody, harmony, rhythm, dynamics, tone color, expressivity, texture, and form. These findings were supported by several empirical studies and experimental results [9], [10]. The superiority of a multifaceted approach was also shown in [11], [12].

In this paper, we propose a multimodal multifaceted MER method for piano music that fuses various kinds of musically-meaningful symbolic, and acoustic features separately extracted from MIDI and audio data, respectively (Fig. 1). Using the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

findings of musical psychology for the network design, a deep neural network (DNN) is expected to be trained effectively with a limited amount of multimodal data. Specifically, our method separately extracts affective features regarding particularly important five factors, i.e., rhythm, dynamics, melody, harmony, and tone color. At the heart of our method is to use a self-attention mechanism that can learn the complicated relationships between different factors and modalities for integrating the extracted features in a weighted manner depending on the musical piece. We show that the proposed approach can achieve the state-of-the-art performance of V/A estimation on the public dataset called EMOPIA [13].

II. Music Emotion Revisited

This section describes five major factors that are considered to play key roles in emotional expression in piano music and which aspects of emotional content are affected. Our theory is based on the findings of recent MER studies and the knowledge in music psychology.

Rhythm refers to the temporal organization and pattern of notes and rests. It determines the timing and duration of each note, creating a sense of pulse, meter, and musical flow. Rhythmic information is usually described as beat, tempo, or accentuation. Rhythm is considered one of the most important elements in emotional expression, it is mainly associated with both the excitement and the happiness of the music.

Dynamics refers to the variations in volume and intensity. It involves the manipulation of loudness and softness to create expressive and emotional effects. It adds depth, contrast, and nuance to the musical interpretation, allowing the pianist to shape the phrases, highlight melodic lines, and create a dramatic impact. Dynamics is also related to the power and intensity of the music and the variation of dynamics may influence the positiveness of the music.

Melody refers to the main musical line or theme that carries the expressive and melodic content of the composition. The prominent sequence of notes stands out and captures the listener's attention. The melody is usually played in the right hand and encompasses the memorable aspect of the music. It carries the emotional and narrative essence of the piece, often conveying a sense of beauty, expression, and storytelling. The melodic elements therefore influence the valence and arousal scores of the emotional expression.

Harmony refers to the simultaneous sounding of multiple notes to create chords and chord progressions. It involves the vertical aspect of music, where different pitches are combined to form harmonious sounds. Harmony provides the foundation and support for the melody, adding depth, richness, and complexity to the overall musical texture. It establishes the tonal framework and creates a sense of stability, tension, and resolution. Harmonic arrangement significantly contributes to the mood, character, and emotional impact of the piece. Major chords are usually related to positive emotions, while minor chords to negative emotions. The complexity of harmony can also be related to the tension and instability of the emotional expression.

Tone color or timbre refers to the unique character of a sound. It is a distinct combination of harmonics, overtones, and how sound is shaped and perceived. Tone color adds richness, variety, and individuality to the musical expression. It can vary according to the specific piano used, the technique employed by the pianist, and the musical context. For example, playing the same note on different parts of the piano keyboard or using different levels of touch can produce different tone colors. Tone color plays a significant role in conveying emotions, creating contrasts, and highlighting musical ideas. Brighter tone color is related to positive or exciting emotions.

Based on the research of Renato et al. [8], we selected some music features that are most suitable for representing the five major factors and designed our model for interpretable piano music emotion recognition.

III. Proposed Method

This section describes the proposed MER method that can consider the multifaceted and multimodal nature of music emotion (Fig. 2). The key feature of the proposed method is to separately analyze the major factors that form the emotion and integrate them in a data-adaptive manner. Our method is implemented on a standard multi-branch DNN consisting of three separate feature extraction branches, the feature fusion module, and the classifier.

A. Melody, Rhythm, and Dynamics Analysis for MIDI Data

We assume that the melody, rhythm, and dynamics factors are originally encapsulated in the symbolic domain (musical score), where these factors have close connections to the distributions of note pitches, locations, and velocities, respectively. Most existing MER methods, however, have dealt with only audio data, where such score-originated factors are treated together with audio-originated factors in a mixed way. In this study, we aim to explicitly learn the statistical patterns of notes from musical scores in a musically-meaningful manner faithful to the findings of music psychology.

As a score representation that plays a crucial role in MER, we use the standard MIDI-like representation. A musical score is serialized as a token sequence, where each note consists of the "note-on", "note-off", and "velocity" tokens and the "time shift" token represents the time interval between notes.

To learn contextual relationships and structural information from MIDI data, we use a recurrent neural network (RNN). Because symbolic music datasets with emotion are severely limited in size, we use as a relatively simpler structured model the bidirectional gated recurrent unit (Bi-GRU) instead of the bidirectional long short-term memory (Bi-LSTM) network to avoid overfitting. Our model consists of two stacked units with 256 hidden layer units each. This computation process can be simply described as follows:

$$\overrightarrow{h_t} = \mathrm{GRU}(\mathrm{token}_t, \overrightarrow{h}_{t-1}), \tag{1}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(\mathrm{token}_t, \overleftarrow{h}_{t+1}), \tag{2}$$

$$h_t = \mathrm{Concat}(\overrightarrow{h_t}, \overleftarrow{h_t}). \tag{3}$$

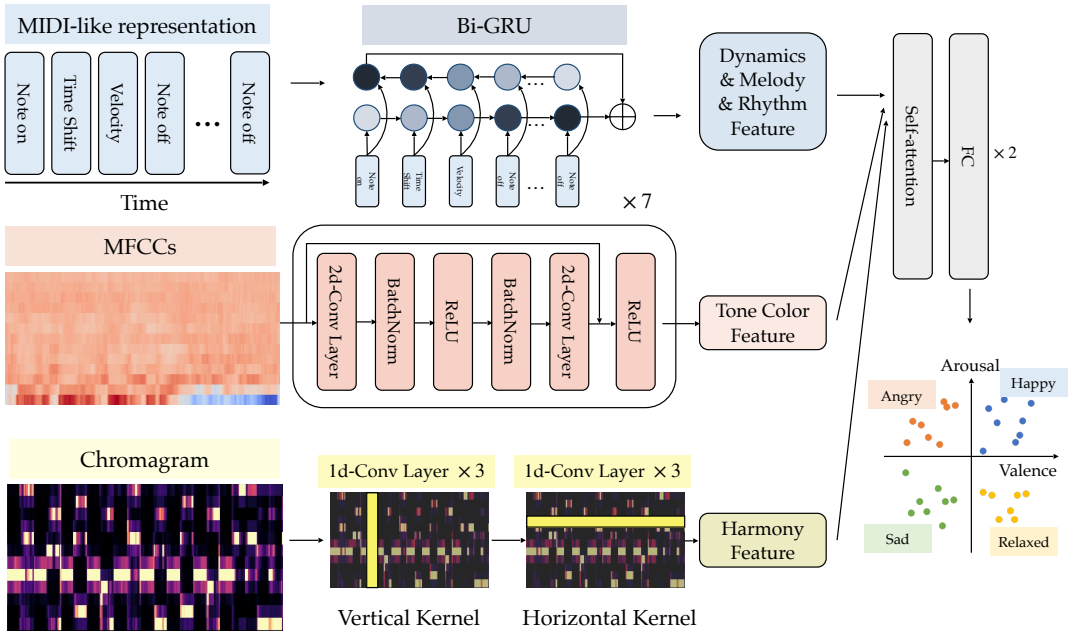

Fig. 2. The proposed network architecture consisting of symbolic and acoustic feature extraction modules and an adaptive feature fusion module with a self-attention mechanism for multimodal multifaceted music emotion recognition.

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ refer to the forward and backward hidden states at time step t respectively, token_t refers to the input MIDI-like token at time step t, h_t is the output hidden state of the Bi-GRU at time step t, which is obtained by concatenating the hidden states in both directions.

After being processed by the trained Bi-GRU, the MIDI-like token sequence is encoded into symbolic domain features encompassing melody, rhythm, and dynamics information. These features can be directly fed into a fully connected classifier for emotion prediction.

B. Harmony Analysis for Chromagrams

We assume that the harmony factor can be effectively extracted from the chromagram of a music signal, which represents the time-varying distributions of twelve pitch classes ranging from C to B in one octave. The chromagram can be obtained from the short-time Fourier transform (STFT) of a music signal with librosa.

To simultaneously learn harmonic structures over frequency and their progressions over time, we use a convolutional neural network (CNN) with a stack of vertical (frequency-directional) and horizontal (time-directional) convolutions on the chromagram. More specifically, we use three Conv1D layers with a vertical convolution kernel to learn the instantaneous information (pitch combination), followed by three Conv1D layers with a horizontal convolution kernel to aggregate the temporal information. The sizes of the vertical and horizontal kernels are (12, 1) and (1, 48), respectively, in this study. Although the pitch or melody information is extracted from the chromagram as well as the harmony information, the self-attention mechanism in the feature fusion module helps this model recognize the essential part of the feature.

C. Tone Color Analysis for MFCCs

We assume that the tone color factor has a close connection to the mel frequency cepstral coefficients (MFCCs) of a music signal, inspired by many studies on various classification tasks (e.g., genre, instrument, and singer classification) including MER. The MFCCs are obtained by applying the discrete cosine transform (DCT) to a log-scale mel-spectrogram. In general, about the first twenty coefficients are used as acoustic features related to the tone color representing the characteristics of the energy distribution and spectral envelope. In this study, the music signal is resampled at 22.05 kHz and the number of mel filterbanks is set to 23, and the 25-dimensional MFCCs are extracted.

To extract the tone color features from the MFCCs, we use a CNN consisting of seven convolutional blocks as in [14]. Each block consists of two Conv2D layers, followed by a 2D batch normalization (BN) layer and a ReLU layer. The kernel size is set to (3, 3) and the stride is set to 2 for every Conv2D layer. We also use the residual connection for efficient learning.

D. Adaptive Fusion of Symbolic and Acoustic Features

We aim to fuse the symbolic and acoustic features separately exacted from the three modules. Since each feature represents only a partial aspect of music emotion, the relative weights of the three features have a large impact on the performance of MER. The basic approach to feature fusion is to design a loss function for multiple features, where the feature weights are treated as hyperparameters that should be optimized in advance based on prior knowledge or with grid search. The performance of this approach based on the fixed weights, however, is limited because the importance of each factor on the overall emotion of the piece can vary considerably.

To adaptively adjust the feature weights according to the characteristics of the musical piece, we use a multi-head self-attention (MHSA) mechanism. Let F be a concatenated feature given by

$$F = \mathrm{Concat}(F_{\mathrm{sym}}, F_{\mathrm{chroma}}, F_{\mathrm{mfcc}}), \tag{4}$$

where F_{sym}, F_{chroma}, F_{mfcc} denote the outputs of the symbolic, harmony, and tone color analysis modules, respectively. A four-headed self-attention is then computed on F to capture the intricate dependencies within F as follows:

$$\mathrm{MHSA}(F) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_4)W_O, \tag{5}$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{6}$$

where head_i denotes an attention head i, W_O is a linear projection matrix, W_i^Q, W_i^K, and W_i^V denote projection matrices applied to Q, K, and V in the attention head i, respectively, and $\mathrm{Attention}$ is defined as

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

where Q, K, V represent the query, key, and value matrices, respectively, and d_k represents the dimensionality of K. The output of the MHSA is flattened and fed into two fully-connected layers, followed by a sigmoid function that predicts the V/A probabilities.

IV. Evaluation

This section reports a comparative experiment conducted to evaluate the proposed and conventional MER methods.

A. Dataset and Metrics

We used a piano music dataset named EMOPIA [13] that contains 1087 piano clips from 387 songs, along with their corresponding MIDI files and emotion labels. Music recordings were collected from YouTube by their corresponding IDs with the 'youtube-dlp' package. We used the same train-validation-test split presented in [13]. Due to the unavailability of several pieces on YouTube, approximately 5% of the original data could not be included in our experiment.

The EMOPIA dataset was annotated with four-quadrant (4Q) emotion labels, where music emotions are categorized into four classes in terms of V/A levels, i.e., "happy" for high arousal and high valence, "relaxed" for low arousal and high valence, "angry" for high arousal and low valence, and "sad" for low arousal and low valence. We used the 4Q classification accuracy as the primary evaluation metric. We also focused on the binary classification accuracies in the V/A estimation. Note that these metrics indicate the performance for recognizing a single aspect of the emotion.

As MIDI-based baseline methods, we tested the symbolic linear regression model (Symbolic-LR) and the LSTM model with an attention mechanism (LSTM-Attn) [15] on the MIDI-like or REMI representation implemented in [13]. As audio-based baseline methods, we tested the acoustic linear regression model (Acoustic-LR) [13] and the short-chunk ResNet model [14]. These uni-modal methods are all lightweight models trained and evaluated only on the EMOPIA dataset.

TABLE I
COMPARISON OF THE PROPOSED AND BASELINE MER METHODS ON EMOPIA DATASET.

Method	4Q	Arousal	Valence
Ours	**0.692**	0.884	**0.869**
LSTM-Attn [15]+MIDI-like [19]	0.684	0.882	0.833
LSTM-Attn [15]+REMI [20]	0.615	0.890	0.746
Symbolic-LR [13]	0.581	0.849	0.651
Audio-LR [13]	0.523	**0.919**	0.558
Short-chunk ResNet [14], [21]	0.677	0.887	0.704

TABLE II
COMPARISON OF PRE-TRAINED MUSIC LANGUAGE MODELS ON EMOPIA DATASET.

Method	4Q Accuracy
Ours	0.692
MidiBERT [16]	0.634
MT-MidiBERT [16], [17]	0.676
QM-MidiBERT [16], [18]	**0.719**

We also tested a large-scale music language model called MidiBERT-Piano [16] and its variants called MT-MidiBERT [16], [17] and QM-MidiBERT [16], [18]. Such models were pre-trained on large-scale MIDI data. Since the V/A estimation was not evaluated on these models, we only used the 4Q metric for comparative evaluation.

B. Results and Discussions

The performances of the proposed and conventional methods are shown in Table. I and Table. II. To the best of our knowledge, our method is the first to combine acoustic and symbolic analyses on MER for piano music. We thus compared only uni-modal methods.

As shown in Table. I, the proposed method achieved the best score of 0.692 in the 4Q metric while keeping the computational cost comparable with the other methods. In arousal estimation, the proposed method performed comparably with the other methods. Since arousal is mainly attributed to low-level acoustic features (e.g., tempo and intensity), even basic audio-based models such as Audio-LR can recognize it accurately. In valence estimation, the proposed method also achieved the best score of 0.869. We found that audio- and MIDI-based methods tended to be better at arousal and valence estimations, respectively. This tendency is consistent with our hypothesis described in Section 1. Overall, by taking into account all factors, the proposed method achieved a better-balanced performance of MER. As shown in Table. II, the proposed model showed a performance comparable to the state-of-the-art large pre-trained model QM-MidiBERT [18], while keeping the computational cost small.

C. Ablation Study

To evaluate the effectiveness of the proposed multimodal multifaceted approach, we conducted an ablation study, where four ablated versions of the proposed method were made by removing two branches out of the three branches in the feature extraction stage or the MHSA mechanism in the feature fusion stage. As shown in Table. III, the full version performed best. This experimental result was consistent with our hypothesis.

TABLE III
ABLATION STUDY OF THE PROPOSED METHOD ON EMOPIA DATASET.

Method	4Q	Arousal	Valence
Full model	**0.692**	**0.884**	**0.869**
Symbolic analysis branch	0.690	0.833	0.744
Harmony analysis branch	0.514	0.726	0.685
Tone color analysis branch	0.649	0.871	0.711
Without MHSA	0.639	0.858	0.753

Since the music score includes information about rhythm, dynamics, and melody related to both valence and arousal, the symbolic analysis branch performed well in all evaluation metrics. Although harmony plays an essential role in emotional expression, it is hard to recognize the overall emotion of a piece only from its harmony information. The harmony analysis branch thus performed worst in all metrics. The tone color analysis branch showed a promising performance, especially in arousal estimation, because the energy distribution and tone color information obtained from MFCCs significantly contributes to the arousal aspect and the overall emotion. By integrating these three branches, the proposed method performed better, especially in valence estimation. We also confirmed the effectiveness of the MHSA-based adaptive feature fusion. The proposed full model outperformed its ablated version by 1.2 pts in the 4Q metric and achieved better balanced V/A estimation performance.

V. CONCLUSION

In this paper, we proposed a multimodal multifaceted MER method for piano music. Inspired by music psychology, we designed a DNN with three branches that separately analyze five main factors of music emotion, i.e., 1) rhythm, dynamics, and melody, 2) harmony, and 3) tone color, from the score, chromagram, and MFCCs of a piece, respectively. Our model fuses the extracted musically-meaningful symbolic and acoustic features with the MHSA mechanism in a data-adaptive manner. In the comparative experiment using the public dataset EMOPIA, our model achieved the state-of-the-art performance in the 4Q metric and a well-balanced performance in the V/A classification metric with a smaller computational cost. The ablation study showed the effectiveness of the multimodal multifaceted approach and that of the MHSA-based adaptive feature fusion. As a future direction, we plan to explore music embedding for MER, which can be used to design a unified MER model that can better capture multiple interdependent factors.

ACKNOWLEDGMENT

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X.

REFERENCES

[1] Juan Sebastián Gómez Cañón, Estefanía Cano, Tuomas Eerola, Perfecto Herrera, Xiao Hu, Yi-Hsuan Yang, and Emilia Gómez. Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Process. Mag.*, 38(6):106–114, 2021.

[2] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. *IEEE Transactions on Multimedia*, 21(12):3150–3163, 2019.

[3] Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. Transformer-based approach towards music emotion recognition from lyrics. In *European Conference on Information Retrieval*, pages 167–175. Springer, 2021.

[4] Ziyu Wang and Gus Xia. Musebert: Pre-training music representation for music understanding and controllable generation. In *ISMIR*, pages 722–729, 2021.

[5] Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho. 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*, pages 310–317, 2020.

[6] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[7] Petri Laukka, Tuomas Eerola, Nutankumar S Thingujam, Teruo Yamasaki, and Grégory Beller. Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion*, 13(3):434, 2013.

[8] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 2020.

[9] Xin Gu, Yinghua Shen, and Jie Xu. Multimodal emotion recognition in deep learning: A survey. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 77–82. IEEE, 2021.

[10] Shreyan Chowdhury and Gerhard Widmer. On perceived emotion in expressive piano performance: Further experimental evidence for the relevance of mid-level perceptual features. *arXiv preprint arXiv:2107.13231*, 2021.

[11] X Hu, K. Choi, and J. S. Downie. A framework for evaluating multimodal music mood classification. *Journal of the American Society for Information Science*, 68(2):273–285, 2017.

[12] Yahan Yu, Bojie Hu, and Yu Li. GHAN: Graph-based hierarchical aggregation network for text-video retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5547–5557, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *arXiv preprint arXiv:2108.01374*, 2021.

[14] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of CNN-based automatic music tagging models. *arXiv preprint arXiv:2006.00751*, 2020.

[15] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[16] Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, and Yi-Hsuan Yang. MidiBERT-piano: Large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*, 2021.

[17] Jibao Qiu, CL Chen, and Tong Zhang. A novel multi-task learning method for symbolic music emotion recognition. *arXiv preprint arXiv:2201.05782*, 2022.

[18] Zhexu Shen, Liang Yang, Zhihan Yang, and Hongfei Lin. More than simply masking: Exploring pre-training strategies for symbolic music understanding. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 540–544, 2023.

[19] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4):955–967, 2020.

[20] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1180–1188, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.