DUAL ATTENTION-BASED MULTI-SCALE FEATURE FUSION APPROACH FOR DYNAMIC MUSIC EMOTION RECOGNITION

Liyue Zhang[1] **Xinyu Yang**[2] **Yichi Zhang**[2]
Jing Luo[2]
[1]The School of Software, Xi'an Jiaotong University , China
[2]The School of Computer Science and Technology, Xi'an Jiaotong University , China
{3121358019, datasonezyc, luojingl}@stu.xjtu.edu.cn, yxyphd@mail.xjtu.edu.cn

ABSTRACT

Music Emotion Recognition (MER) refers to automatically extracting emotional information from music and predicting its perceived emotions, and it has social and psychological applications. This paper proposes a Dual Attention-based Multi-scale Feature Fusion (DAMFF) method and a newly developed dataset named MER1101 for Dynamic Music Emotion Recognition (DMER). Specifically, multi-scale features are first extracted from the log Mel-spectrogram by multiple parallel convolutional blocks. Then, a Dual Attention Feature Fusion (DAFF) module is utilized to achieve multi-scale context fusion and capture emotion-critical features in both spatial and channel dimensions. Finally, a BiLSTM-based sequence learning model is employed for dynamic music emotion prediction. To enrich existing music emotion datasets, we developed a high-quality dataset, MER1101, which has a balanced emotional distribution, covering over 10 genres, at least four languages, and more than a thousand song snippets. We demonstrate the effectiveness of our proposed DAMFF approach on both the developed MER1101 dataset, as well as on the established DEAM2015 dataset. Compared with other models, our model achieves a higher Consistency Correlation Coefficient (CCC), and has strong predictive power in arousal with comparable results in valence.

1. INTRODUCTION

With the rising demand for music consumption and the explosive growth of music content, Music Emotion Recognition (MER) demonstrates its critical position in music understanding and applications. It has been widely used in personalized music recommendation [1], music therapy [2], music education [3], music generation [4], etc.

To portray human emotions, two main types of models were differentiated in the past [5]: discrete emotion model [6, 7] and dimensional emotion model [8–11]. The discrete emotion model describes human emotion as categorical adjectives, such as happiness, anger, sadness, joy, etc. However, limited words cannot adequately describe human emotions, different emotions are better described on a continuous scale than as a set of discrete values. In Russell's two-dimensional valence-arousal (V-A) emotional model [12], emotions are described as points on the plane that is spanned by the arousal and valence axes. This turns the problem of emotion prediction into a two-dimensional regression issue based on Russell's emotion model. This paper is focused on the study of Dynamic Music Emotion Recognition (DMER), which predicts the emotion of music using continuous V-A values at a short interval.

Among the existing studies, Long Short-Term Memory (LSTM) has received extensive attention in the DMER due to its superiority in sequence modeling [8, 13–15]. Convolutional Neural Network (CNN) is used to extract features in many fields. Researchers have recently focused on improving emotion recognition accuracy using a combination of CNN and Recurrent Neural Network (RNN) [9, 16–18]. However, LSTM-based models still use handcrafted features as input, and some widely used handcrafted feature operations will lose high-level features. The CNN-RNN-based model mainly uses a fixed-scale CNN. Due to its fixed receptive field, the learned CNN features are limited, and the emotional crucial features of different fields of view are not extracted. Moreover, various problems exist in existing music emotion datasets, which also hinder the progress of DMER.

This paper proposes a novel Dual Attention-based Multi-scale Feature Fusion (DAMFF) model and develops the music emotion dataset MER1101 for DMER. On the one hand, our model first utilizes multi-scale convolution to extract features at different temporal-frequency spans from the log Mel-spectrogram. Then, we propose a Dual Attention Feature Fusion (DAFF) module for fusing multi-scale context features from spatial and channel dimensions to enhance the expressive ability of CNN. Finally, the BiLSTM model processes these features and predicts V-A emotional labels. On the other hand, we develop a high-quality dataset named MER1101. Compared with the existing publicly available datasets in the MER domain, MER1101 contains 1101 music snippets from 16 genres with richer languages, more extensive size, and more balanced emotion label distribution. We evaluate our method using the MER1101 dataset and DEAM2015 [19] dataset.

Attribution: L. Zhang, X. Yang, Y. Zhang, J. Luo, "Dual Attention-based Multi-scale Feature Fusion Approach for Dynamic Music Emotion Recognition", in *Proc. of the 24th Int. Society for Music Information Retrieval Conf.,* Milan, Italy, 2023.

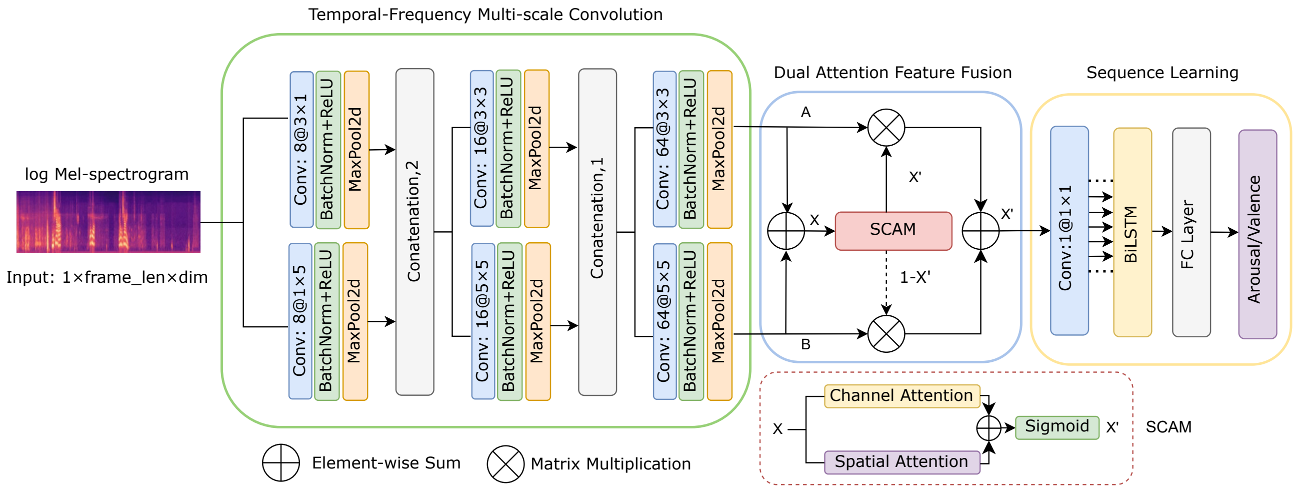


Figure 1. DAMFF model architecture. The input is 2D spectrogram. The architecture combines temporal-frequency multi-scale feature extraction, dual attention feature fusion, and sequence learning to achieve dynamic emotion prediction for music.

On the MER1101 dataset, we achieve a Consistency Correlation Coefficient (CCC) of 0.4223 for arousal and 0.1115 for valence. On the DEAM2015 dataset, we achieve a C-CC of 0.4203 for arousal and 0.0151 for valence. Experimental results show our method outperforming a number of baseline and SOTA models in DMER, by means of an improved CCC metric.

2. RELATED WORK

Researchers have made many efforts in the DMER in the past few years. In the early days, RNN made a breakthrough in this field due to their advantages in sequence processing. In the "Emotion in Music" task at MediaEval from 2013 to 2015, LSTM-based methods achieved state-of-the-art performance [20]. Li *et al.* [8] pointed out that in music composition, performance, and annotation, the emotion in music is related to the previous and future contexts. Therefore, they chose Bidirectional LSTM (BiLSTM) as the regression model and proposed a multi-scale fusion method based on an Extreme Learning Machine (ELM) to improve the performance of the BiLSTM model. But the LSTM-based models mentioned above use suboptimal hand-crafted features as input, making it difficult to improve emotion recognition.

Later, researchers began to employ CNN for high-level invariant features extracted from raw music data [21–23]. Pons *et al.* [24] discussed how convolution filters with different shapes are suitable for specific musical concepts and experimentally proved that the size of CNN filters can be interpreted in both the temporal and frequency dimensions of the spectrogram. Researchers have combined CNN and RNN to improve the accuracy of emotion recognition, Malik *et al.* [16] proposed a two-dimensional V-A space continuous emotion prediction method composed of stacked convolution and recurrent neural network. Compared to using BiLSTM [15] only, this method achieved better results with fewer parameters; Dong *et al.* [9] replaced the connection between the input layer and the hidden layer of the RNN with a CNN to adaptively learn the sequential-information-included affect-salient features from the spectrogram; Zhang *et al.* [25] extracted MFCCs and Cochleagrams from raw music data as input features, and adopted an audio feature fusion method based on the combination of CNN and BiLSTM to predict the emotional V-A values in music. However, CNN-RNN-based models still have problems with limited convolutional receptive fields. For MER, due to the limited size of the convolution kernel, the convolution is mainly biased towards learning local information, which is insufficient for learning the correlation between the spatial and channel axes.

Various attention mechanisms are devised to solve the above problem in speech emotion recognition [21, 26, 27]. Guo *et al.* [26] proposed a representation learning method with spectral-temporal channel (STC) attention, which was integrated with CNN to improve representation learning ability; Zhang *et al.* [21] applied multi-scale region attention in deep convolutional neural networks to focus on emotional features at different granularities; Zhang *et al.* [27] implemented an attention layer on the arousal, valence, and dominance tasks and completed multi-task predictions to capture the contribution of different parts of each task. Nonetheless, the attention mechanism is currently not widely applied in the field of DMER.

In this paper, we propose a novel attention module, the Spatial Channel Attention Module (SCAM), which considers spatial and channel dimensions to capture the relative importance of features and integrates multi-scale convolutions for enhanced representation learning. We aim to build an attention mechanism that extracts salient information from multiple dimensions and can fuse contextual information.

J. S. Gómez-Cañón et al. [28] summarized existing MER datasets. But they have some problems, for example, some datasets have insufficient number of music, and some datasets have no dimension labels. After our comprehensive comparison, the three datasets CH818 [29], PMEmo [30] and DEAM [19] are relatively suitable for the DMER task. However, all three datasets have some disadvantages. The songs in the CH818 dataset only contain Chinese pop songs and are not public, while PMEmo only Western pop songs; The annotators and annotating times of the training set and evaluation set in the DEAM2015 dataset are different, resulting in a discrepancy in performance [19]. To enrich existing musical emotion datasets, we develop a high-quality dataset, MER1101. MER1101 contains 1101 music snippets, which is better than most datasets in the MER domain in terms of genre, language, number of music, and has more balanced distributed emotion annotations.

3. METHODOLOGY

The proposed DMER processing method consists of three phases. Firstly, we build a Temporal-Frequency Multi-scale Convolution network using three different shapes of convolutional filters. Secondly, we propose a Dual Attention Feature Fusion network to focus more on the channel and spatial with important information and fuse multi-scale convolutional features in different dimensions. And finally, we employ BiLTSM, building a map from emotion-crucial features to emotional space. The specifics are as follows.

3.1 Temporal-Frequency Multi-scale Convolution

CNN has been proven effective at tackling various visual tasks [31, 32]. In vision tasks, the filter dimension has spatial meaning, and the audio spectrogram filter dimension corresponds to temporal and frequency [24]. We design a temporal-frequency multi-scale convolution module with three types of filters to capture various musical features. From the musical point of view, the temporal filter (*1-by-n*) can learn temporal dependence in music; the frequency filter (*m-by-1*) can learn pitch and timbre, and the square filter (*m-by-n*) can learn different musical features according to the size of *m* and *n*. As shown in Figure 1, we extract features through three layers of parallel convolutional blocks in the Temporal-Frequency Multi-scale Convolution module.

Firstly, we take the 30-second log Mel-spectrogram as input and perform distinct convolution operations on each 0.5-second segment to keep the individual properties at each moment. Secondly, the first layer introduces 3×1 and 1×5 filters to capture features along the temporal and frequency axes, and their outputs are concatenated along the time dimension. Finally, the concatenated results of the first layer are put into consecutive parallel convolutional layers with kernel sizes 3×3 and 5×5. The output of the second layer is concatenated along the channel dimension, while the output of the third layer is fed into a dual attention feature fusion module for feature fusion. After each convolutional layer, batch normalization [33], the ReLU function [34] and max pooling are applied.

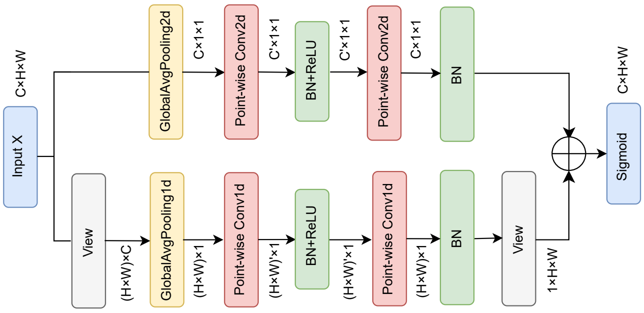


Figure 2. SCAM model architecture.

3.2 Dual Attention Feature Fusion

To further enhance the representation ability of CNN and capture the important information, we design a Dual Attention Feature Fusion (DAFF) module to focus more on the channel and spatial with important information for fusing multi-scale convolutional features. As shown in Figure 1, the DAFF module includes the Spatial Channel Attention Module (SCAM). By element-wise summing the outputs of 3×3 and 5×5 convolutions, we get a feature map $X \in R^{C\times H\times W}$ as input to SCAM, which is then fed into the spatial and channel attention modules, respectively. In Sections 3.2.1 and 3.2.2, we describe the proposed SCAM in detail.

3.2.1 Channel Attention Module

We convert a single channel into 64 channels through the Temporal-Frequency Multi-scale Convolution, strengthening the temporal correlation between distinct channels. In this case, we use the channel attention mechanism, which focuses on *what* the essential features are. While traditional attention mechanisms only focus on temporal structures, channel attention can learn the importance of different channels to deactivate features that do not contribute much to emotion. Figure 2 shows the channel attention module, similar to the Squeeze-and-Excitation block [35]. The module is mainly divided into two parts: squeeze and excitation operations. Specifically, given an input feature $X \in R^{C\times H\times W}$, we first use Global Average Pooling independently for each channel to aggregate spatial information and generate a channel attention map $C \in R^{C\times 1\times 1}$. Next, we perform an excitation operation using two pointwise convolutions to enable cross-channel interaction. Finally, the channel attention map $C \in R^{C\times 1\times 1}$ is obtained. In short, the channel attention map is calculated as follows:

$$C = \beta(Conv2d_2(\delta(\beta(Conv2d_1(Pool2d(X)))))) \quad (1)$$

where δ and β denote the ReLU function and batch normalization, respectively, and $Pool2d$ and $Conv2d$ represent the global average pooling2d and point-wise convolution2d, respectively.

3.2.2 Spatial Attention Module

We propose the spatial attention model, which exploits the spatial relationship between features to generate a spatial attention map. Spatial attention focuses on *where* the important features are and supplements channel attention.

The spatial attention module obtains the spatial attention map in four phases. First, the input feature through view operation is converted into a spatial feature map $S' \in R^{(H \times W) \times C}$. Second, a global pooling operation is applied along the channel axis to compress the channels to obtain spatial-level features. Third, we use two point-wise convolutions to execute excitation operations and get feature weights at distinct positions. Finally, the resulting spatial attention map is translated into $S \in R^{1 \times H \times W}$. In short, the spatial attention map is calculated as follows:

$$S = \beta(Conv1d_2(\delta(\beta(Conv1d_1(Pool1d(X)))))) \quad (2)$$

where δ and β denote the ReLU function and batch normalization, respectively, and $Pool1d$ and $Conv1d$ represent the global average pooling1d and point-wise convolution1d, respectively.

After that, we perform an element-wise sum operation on the output of the dual attention and through the sigmoid function to obtain a new attention weight map $X' \in R^{H \times W \times C}$.

$$X' = Sigmoid(S \oplus C) \quad (3)$$

3.2.3 *Feature Fusion Strategy*

In order to effectively aggregate multi-scale context information, we introduce the fusion strategy in [36], as shown by Dual Attention Feature Fusion in Figure 1. The output of SCAM is represented as X', $1 - X'$ by the solid line and dotted line, respectively. Based on the SCAM, the multi-scale feature fusion can be expressed as:

$$Z = X' \otimes A + (1 - X') \otimes B \quad (4)$$

where A and B represent the outputs of 3×3 and 5×5 convolutions respectively, $Z \in R^{C \times H \times W}$ is the fused feature.

3.3 Sequence Learning

Through the DAFF module, we get emotion-crucial features from multi-scale convolutional features. After reducing dimension, the features of the entire 30s of music snippet are input into the Bidirectional LSTM (BiLSTM) for long-term sequence learning. Finally, the emotional features are mapped to the emotional space with the help of a fully connected layer.

4. EXPERIMENTS

4.1 Dataset

We conduct our experiments on the DEAM2015 [19] dataset and our newly developed dataset MER1101. The details of each dataset are given below.

DEAM: This dataset was developed in the “Emotion in Music” (EiM) task [37] of the MediaEval benchmark. We utilized the DEAM2015 dataset, with the training set consisting of 431 30-second samples and the evaluation set consisting of 58 full-length songs. This dataset is the most commonly used benchmark in dynamic music emotion recognition, but Cronbach’s α of the evaluation set is 0.29±0.94 for valence, which is relatively low [19]. Furthermore, due to the different spatio-temporal environments and annotators of the emotion annotation process of the training set and the evaluation set [19], the performance derived from the training and evaluation set shows a non-negligible discrepancy, especially in the valence dimension.

MER1101 [1] **:** Similar with DEAM, MER1101 is also based on Russell’s valence-arousal emotion model. It contains 1101 music snippets gathered from the internet, with each ranging in duration from 16.5 seconds to 125.5 seconds. The dataset has both discrete and dimensional labels. Every song in the dataset has been annotated by three music experts and ten college students. The annotators listened to the song once and annotated the emotional adjectives of the song. After they were familiar with the song, they listened to it twice and annotated the V-A values. Annotators were only paid the full fee after their work had been reviewed. Student-labeled Cronbach’s α arousal is 0.6295 ± 0.3574, 0.5624 ± 0.3766 for the valence. Expert-labeled Cronbach’s α arousal is 0.3556 ± 0.3442, 0.2420 ± 0.3148 for the valence.

Compared with other music datasets, the MER1101 dataset has the following four advantages: 1) The dataset contains more genres (16 genres), including pop, DJ dance, chinoiserie, electronic, hip-hop, etc.; 2) It contains richer language, meeting the ratio of nearly 5:3:1:1 for Chinese, English, Japanese and Korean, and other languages; 3) The samples in our dataset distribute more balanced in the emotional quadrants and there are no more than three songs by the same artist in each V-A quadrant; 4) The size of our dataset is relatively larger than the current music datasets.

Our dataset can be used for a variety of music tasks, such as music genre classification, music generation with emotion, music emotion recognition, *etc*.

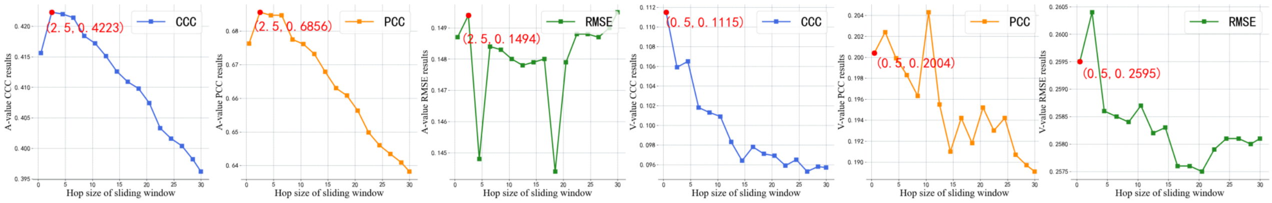
4.2 Evaluation Metrics

We use the Concordance Correlation Coefficient (CCC), Pearson Correlation Coefficient (PCC), and Root Mean Square Error (RMSE) as evaluation metrics. Each metric is computed by the ground-truth and predicted V-A values for each song and averaged across songs. The CCC combines the characteristics of PCC and RMSE to evaluate not only the trend of emotional changes but also the disparity between predictions and ground-truth. As a result, we consider CCC to be the most important evaluation metric.

4.3 Experimental Setup

Since DEAM2015 predefines the training and evaluation set configuration, we only describe the dataset division for MER1101 here. Firstly, we choose 925 songs lasting more than 30 seconds from the MER1101 dataset and randomly split them into a training set (80% of the data) and an evaluation set (20% of the data). Then, we split each song in

[1] See https://ismir-2023.github.io/MER1101/ for details.



(a) A-value CCC (b) A-value PCC (c) A-value RMSE (d) V-value CCC (e) V-value PCC (f) V-value RMSE

Figure 3. The CCC, PCC, and RMSE of arousal and valence with different hop sizes on the MER1101 dataset.

the training set into 30-second segments and kept complete songs for the evaluation set. The final training set contains 1526 30-second music snippets, and the evaluation set contains 185 complete songs. The DEAM dataset uses the official training and evaluation sets. To obtain a more accurate comparison and minimize accidental errors, we use 5-fold cross-validation on both datasets.

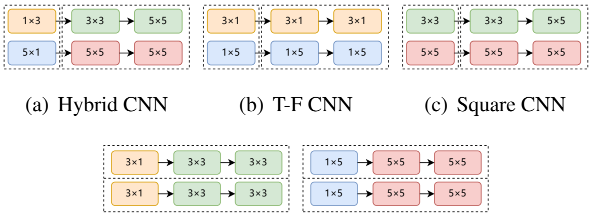
The log Mel-spectrogram is extracted using librosa [38], where the Mel band is 128, the sampling rate is 44100Hz, and the window size and hop size are 60 ms and 10 ms, respectively. The size of the convolution kernel is shown in Figure 1. We utilize the Adam optimizer for training, with learning rate of 0.0003, training epoch of 100, and batch size of 32. To prevent overfitting, we adopt the early stopping strategy. In addition, we use CCC and RMSE as loss functions for arousal and valence, respectively.

4.4 Experimental Results

4.4.1 Hop Size Selection of Sliding Window

For the MER1101 dataset, we train the model with music snippets of fixed duration, while the durations of the music snippets are variable during the test. Thus, we could not directly predict the emotion of the whole music. We propose a sliding window-based testing scheme to address this issue and ensure the continuity of the predicted V-A curves. During testing, we utilize the window size of T seconds and the hop size of t seconds. Each T second of audio in the window is input to the model, and the corresponding T seconds V-A curves are predicted. The first window takes the prediction result of T seconds, and each subsequent window only takes the result of the last t seconds.

We investigate the impact of hop size on the results of music emotion recognition on the MER1101 dataset. We set the window size to 30s, the same as the training set sample duration. Figure 3 shows the experimental results, CCC and PCC change significantly and show a downward trend with increasing hop size, and the change in RMSE is not obvious. We observe that with the increase of the hop size, the emotion prediction effect decreased significantly, demonstrating that the shorter hop size performs better. During listening to music, the user's emotion at a certain moment is an accumulation of previous music content. Therefore, providing the model with as much context as possible benefits emotion recognition. A shorter hop size can provide more context for the model to predict the current musical mood. In the experiments on the MER1101 dataset below, we adopt hop sizes of 2.5s and 0.5s for arousal and valence, respectively.



(a) Hybrid CNN (b) T-F CNN (c) Square CNN

(d) T-S CNN (e) F-S CNN

Figure 4. Five CNN architectures.

Model	Arousal			Valence		
	CCC ↑	PCC* ↑	RMSE* ↓	CCC ↑	PCC ↑	RMSE ↓
Hybrid CNN	**0.4223**	0.6856	0.1494	**0.1115**	**0.2004**	0.2595
T-F CNN	0.4120	0.6787	0.1478	0.0846	0.1363	0.2684
Square CNN	0.4130	0.6894	0.1439	0.0732	0.1343	0.2703
T-S CNN	0.4090	0.6881	0.1458	0.1085	0.1959	**0.2542**
F-S CNN	0.4150	0.6804	0.1562	0.1046	0.1640	0.2800

* The result of the significance test (Student's t test) show that there is no significant difference between the results of this metric.

Table 1. Experimental results of different CNN architectures on the MER1101 dataset.

4.4.2 Impact of CNN filters

In this section, we compare the influence of different CNN architectures on the experimental results of the MER1101 dataset. In this paper, we adapt three types of convolution: temporal filters (*1-by-n*), frequency filters (*m-by-1*), and squared filters (*m-by-n*). Convolution filters of different shapes have different musical concepts. We combined them into five architectures. In Figure 4(a), the CNN architecture used here is a "Hybrid CNN" architecture. Figure 4(b) uses the temporal filters and frequency filters, and we call it the "T-F CNN" architecture. Figure 4(c) only uses a square filter, so we call it "Square CNN" architecture. Figure 4(d) and Figure 4(e) are referred to as "T-S CNN" and "F-S CNN", respectively. The experimental results are shown in Table 1, which show that the "Hybrid CNN" architecture has better expressiveness on the DMER

Model	MER1101 dataset						DEAM2015 dataset					
	Arousal			Valence			Arousal			Valence		
	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC↑	RMSE ↓	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC↑	RMSE ↓
CRNN [16]	0.2798	0.5177	0.1625	0.0573	0.1033	0.2721	0.3488	0.5885	**0.2197**	0.0053	-0.0292	0.3542
BCRSN [9]	0.1741	0.3770	0.3063	0.0660	-0.0647	0.4143	0.3168	0.5148	0.2397	0.0125	-0.0171	**0.2914**
DNN [17]	0.0529	0.0903	0.2372	0.0118	0.0017	0.2734	0.2757	0.4282	0.2483	0.0075	0.0031	0.3353
MCRNN [18]	0.0564	0.0918	0.2401	0.0155	0.0028	0.2752	0.2700	0.4396	0.2428	0.0137	0.0126	0.3135
DAMFF	**0.4223**	**0.6856**	**0.1494**	**0.1115**	**0.2004**	**0.2595**	**0.4203**	**0.6866**	0.2401	**0.0151**	**0.0366**	0.3403

Table 2. Compared with the existing results.

task. It is shown that extracting them simultaneously is beneficial to obtain music emotion information from different perspectives, and the PCC and RMSE changes of Arousal are not significant.

4.4.3 *Comparison with the Existing Models*

We compare the DAMFF to other DMER methods [9, 16–18] published in recent years. They differ from us in that [18] takes DEAM2014 [39] as the dataset, which consists of 744 songs. [16–18] take RMSE as the evaluation metrics, and [9] translates numerical-type V-A values to binary representation and independently predict emotion for each 0.5s.

In this paper, we reproduce the models mentioned above on the DEAM2015 and MER1101 datasets. All models' performance is evaluated with the same experimental configurations, i.e., the same dataset, evaluation metrics, and metric calculation method. Table 2 shows the results of the experiments. On the MER1101 dataset, our model is superior to the others in all three metrics. On the DEAM2015 dataset, our model shows powerful recognition ability for arousal, but the valence slightly outperforms the previous models, which may stem from the less consistent valence annotations [15]. We believe predicted valence values on the DEAM2015 dataset are relatively incapable of evaluating DMER since the predicted CCC value number in valence driven from all models is near zero. Experiments show that our model can perform well in emotion recognition on different datasets, especially in the arousal dimension. Overall, valence values are more impoverished in both datasets than arousal values, indicating that predicting valence is more challenging. This is also consistent with the conclusions of most works.

Model	Arousal			Valence		
	CCC ↑	PCC* ↑	RMSE* ↓	CCC ↑	PCC ↑	RMSE ↓
DAMFF	**0.4223**	0.6856	0.1494	**0.1115**	**0.2004**	**0.2595**
w/o Fusion Strategy	0.4097	0.6869	0.1563	0.1074	0.1722	0.2707
w/o Channel Attention	0.4061	0.6740	0.1494	0.1071	0.1904	0.2650
w/o Spatial Attention	0.4177	0.6819	0.1518	0.1009	0.1874	0.2720
w/o DAFF	0.3982	0.6813	0.1562	0.0977	0.1693	0.2670

* The result of the significance test (Student's t test) show that there is no significant difference between the results of this metric.

Table 3. Ablation experiments of arousal and valence on the MER1101 dataset.

4.4.4 *Ablation Study*

To investigate the role of various modules, we constructed four ablation modules. Among them, "w/o Fusion Strategy" directly inputs the result of the SCAM module into BiLSTM, which explores the role of fusion strategy. In addition, the influence of dual attention is studied using "w/o Channel Attention", "w/o Spatial Attention", and "w/o DAFF". Table 3 shows the experimental results on the MER1011 datasets. The results show that: 1) the non-linear fusion strategy of the attention mechanism better aggregates the multi-scale context and performs better; 2) the attention mechanism increases the weights of emotional features, which is helpful for emotion recognition. At the same time, dual attention is better than single attention, indicating that spatial and channel attention mechanisms learn and emphasize *what* and *where* affect-salient features, effectively improving CNN features. In summary, we conclude that fusing multi-scale convolutional features from spatial and channel dimensions is more conducive to capturing key emotional features, which is more evident on the CCC metric.

5. CONCLUSION

This paper proposes a novel Dual Attention-based Multi-scale Feature Fusion (DAMFF) network, which extracts multi-scale convolutional features from spectrograms and exploits the dual-attention mechanism to capture important channel and spatial information. The network adopts the fusion mechanism that aggregates multi-scale context information, effectively improving CNN features' expressive ability. The music emotion dataset MER1101 we developed contains 1101 music audio with 16 genres, 5 languages and a balanced distribution of emotion labels. Experimental results show that our model outperforms the comparison methods on the CCC metric on both MER1101 and DEAM2015 datasets. Furthermore, our model has substantial prediction capabilities in terms of arousal and comparable results in terms of valence.

The prediction of the valence dimension is still challenging in DMER. In the future, we will focus on developing more effective techniques, such as pre-training audio features for improving the recognition performance of valence.

6. REFERENCES

[1] S. M. Florence and M. Uma, "Emotional detection and music recommendation system based on user facial expression," in *IOP Conference Series: Materials Science and Engineering*, vol. 912, no. 6. IOP Publishing, 2020, p. 062007.

[2] G. A. Dingle, P. J. Kelly, L. M. Flynn, and F. A. Baker, "The influence of music on emotions and cravings in clients in addiction treatment: A study of two clinical samples," *The Arts in Psychotherapy*, vol. 45, pp. 18–25, 2015.

[3] T. Xia, Z. Li *et al.*, "Behavioral training of high-functioning autistic children by music education of occupational therapy," *Occupational Therapy International*, vol. 2022, 2022.

[4] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801*, 2020.

[5] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia systems*, vol. 24, pp. 365–389, 2018.

[6] K. Trohidis, G. Tsoumakas, G. Kalliris, I. P. Vlahavas *et al.*, "Multi-label classification of music into emotions." in *International Conference on Music Information Retrieval (ISMIR)*, vol. 8, 2008, pp. 325–330.

[7] J.-H. Su, T.-P. Hong, Y.-H. Hsieh, and S.-M. Li, "Effective music emotion recognition by segment-based progressive learning," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 3072–3076.

[8] X. Li, H. Xianyu, J. Tian, W. Chen *et al.*, "A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 544–548.

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

[10] S. Chaki, P. Doshi, S. Bhattacharya, and P. Patnaik, "Explaining perceived emotion predictions in music: An attentive approach." in *International Conference on Music Information Retrieval (ISMIR)*, 2020, pp. 150–156.

[11] Z. Huang, S. Ji, Z. Hu, C. Cai, J. Luo, and X. Yang, "ADFF: Attention Based Deep Feature Fusion Approach for Music Emotion Recognition," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, 2022, pp. 4152–4156.

[12] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[13] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 5412–5416.

[14] Y. Ma, X. Li, M. Xu, J. Jia, and L. Cai, "Multi-scale context based attention for dynamic music emotion prediction," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1443–1450.

[15] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai, "Dblstm-based multi-scale fusion for dynamic emotion prediction in music," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[16] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Proceedings. 14th Sound Music Comput. Conf.*, 2017, pp. 208–213.

[17] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "Dnn based music emotion recognition from raw audio signal," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–4.

[18] N. He and S. Ferguson, "Multi-view neural network-s for raw audio-based music emotion recognition," in *2020 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2020, pp. 168–172.

[19] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2015," in *MediaEval*, 2015.

[20] Aljanaki, Anna and Yang, Yi-Hsuan and Soleymani, Mohammad, "Developing a benchmark for emotional analysis of music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[21] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech e-motion recognition with multiscale area attention and data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6319–6323.

[22] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[23] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 6437–6441.

[24] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *2016 14th international workshop on content-based multimedia indexing (CBMI).* IEEE, 2016, pp. 1–6.

[25] C. Zhang, J. Yu, and Z. Chen, "Music emotion recognition based on combination of multiple features and neural network," in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, vol. 4. IEEE, 2021, pp. 1461–1465.

[26] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 6304–6308.

[27] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6705–6709.

[28] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[29] X. Hu and Y.-H. Yang, "The mood of chinese pop music: Representation and recognition," *Journal of the Association for Information Science and Technology*, vol. 68, no. 8, pp. 1899–1910, 2017.

[30] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, 2018, pp. 135–142.

[31] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[32] M. Zhao, G. Cao, X. Huang, and L. Yang, "Hybrid transformer-cnn for real image denoising," *IEEE Signal Processing Letters*, vol. 29, pp. 1252–1256, 2022.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning.* pmlr, 2015, pp. 448–456.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[36] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3560–3569.

[37] "Mediaeval benchmarking initiative for multimedia e-valuation," http://www.multimediaeval.org/.

[38] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[39] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013, pp. 1–6.