ICFF-Net: Interlaced cross-attention feature fusion network for music genre classification

Shiting Meng, Cairui Yan, Yingyuan Xiao✉, Wenguang Zheng, and Xu Cheng✉

Computer Science and Engineering, Tianjin University of Technology, Tianjin, China.
mengsting030@163.com, 18406550401@163.com, yyxiao@tjut.edu.cn,
wenguangz@tjut.edu.cn, xu.cheng@ieee.org

Abstract. Fusion music combines diverse rhythms, melodies, and stylistic elements to create a complex and creative musical form. However, traditional classification methods often focus on improving overall classification accuracy, neglecting the unique nature of fusion music genres, which results in lower accuracy. To accurately classify fusion music, we must deeply consider the interactions between different features and perform detailed feature processing. In this paper, we propose an interlaced cross-attention feature fusion network, called ICFF-Net. The network aims to achieve optimal decision results by leveraging the advantages of multiple features and adjusting the weights assigned to different genres. Specifically, we extract features from the spectrograms (2D) of the original audio and the accompaniment audio (1D). The temporal and frequency domain features of the 1D-features are concatenated with the 2D-features as two sources. We then construct an interlaced cross-attention module with a dual-layer structure (IC-Attention) to fuse these features. IC-Attention employs cross-attention with two embeddings from different sources, comprehensively considering the impact of both 2D and 1D features on the classification results, achieving comprehensive feature interaction and weight allocation. Additionally, we constructed a Fusion-Music dataset to analyze classification bias's effect on fusion genres. Our method demonstrates superior performance in addressing classification bias.

Keywords: Music genre classification · Feature fusion · Cross-attention.

1 Introduction

Music genre classification (MGC) [5] is a widely explored research domain in the field of music information retrieval (MIR), holding significant importance in various aspects such as music recommendation, search engines, and music composition. The diversification of music styles and the fusion of different genres has introduced new challenges in the domain of music genre classification. Fusion music requires a focus on musical features such as rhythm and melody, necessitating more detailed feature processing and consideration of the interplay between multiple features.

The classification task requires objective and distinct categories, yet compared to other classification tasks, music genre labels are more subjective. There is no universally accepted method for defining genres. With the continuous evolution of music, artists continuously blend characteristics and styles from different genres, creating new music genres. This phenomenon, known as genre fusion, results in music exhibiting similar properties across genres. For example, rock, pop, and country music have interwoven histories in music. These three genres exhibit traces of blues music in aspects like melody, rhythm, and harmony. Particularly in the case of rock music, it evolved from blues, country music, and rhythm and blues (R&B). This evolution also led to the formation of a new genre known as country rock, which not only profoundly influenced subsequent country and rock music but also provided inspiration for the fusion and innovation of other music genres. Such genre fusion showcases the diversity and innovation within music, paving new paths for its development. However, with similar musical traits shared between fused genres, traditional classification methods often struggle to categorize fusion genres, resulting in poorer classification performance.

With the successful application of deep learning[7][13], researchers increasingly view MGC tasks as image classification tasks based on audio spectrograms, where the choice of features often determines the classification's effectiveness. Some researchers have considered a variety of features, including time-frequency features (such as spectrograms), temporal features of audio signals, and handcrafted features (such as mel-frequency cepstral coefficients (MFCC), short-time energy [15], rhythm content, and pitch content feature sets [20], etc.). This comprehensive consideration allows for information to be obtained from different perspectives, leading to better decision results. MusicNeXt [17] designed a lightweight CNN network that emphasizes musical features. Employing within-class angle constraints to strengthen learning differences among various types has reduced bias in the classification outcomes; Chang et al. [6] proposed MS-SincNet for learning 2D representations from 1D raw waveform signals while considering both one-dimensional and two-dimensional features for classification; Jaehun Kim et al. [11] extracted artist group feature factors expression, considering different types of artist information for inference in MGC tasks; Allamy et al. [1] constructed a 1D CNN directly learning time-frequency features from raw audio signals and learning multiple filters for classification tasks; Xie et al. [22] proposed a 1D res-gated CNN model to extract local information of audio sequences; Cai et al. [3] proposed an auditory-inspired feature set, merging logarithmic spectral, timbral-based, and psychoacoustic acoustic feature sets for fusion; Zhao [24] introduced a music classification model based on a BP neural network, integrating diverse music features for the classification task. These studies, based on feature extraction and various fusion attempts, have improved classification performance.

Based on the aforementioned research, we found that traditional classification methods when performing classification tasks, overlook the need to handle the more complex and diverse features of fusion music and fail to fully consider the uniqueness of fusion genres, resulting in the underutilization of feature information. Additionally, these methods lack appropriate multi-feature fusion techniques and cannot address the diversity and complexity of fusion music. Detailed weighting of features and balancing the weights of different types of features are crucial for resolving classification bias. To address these issues, this paper introduces an Interlaced Cross-attention Feature Fusion Network (ICFF-Net) aimed at resolving bias in fusion music classification. Specifically, we extract features from the spectrogram of the original audio (2D) and the accompaniment audio (1D) and concatenate the time-domain and frequency-domain features of the 1D features with the 2D features as two feature sources. Subsequently, we construct a dual-layer interlaced cross-attention module (IC-Attention) to fuse these two types of features. IC-Attention employs interlaced cross-attention embeddings from different sources twice, fully considering the impact of 2D and 1D audio features on classification results, thereby achieving comprehensive feature interaction and weight allocation.

Our contributions in this paper can be summarized as follows:

(1)We propose an interlaced cross-attention feature fusion network (ICFF-Net) to address bias in fusion music classification; (2)We extract features from the spectrogram of the original audio (2D) and the accompaniment audio (1D) and perform data augmentation on the accompaniment audio to expand the accompaniment dataset; (3)We construct an interlaced cross-attention module with a dual-layer structure that enriches 2D and 1D features through mutual enhancement, enabling effective feature interaction and fusion; (4)We construct a dataset called FusionMusic to evaluate the model's performance in distinguishing fusion genres, and our method demonstrate superior performance in addressing classification bias in fusion genres.

2 Method

In this section, we introduce the proposed ICFF-Net, which is a musical features enhancement network based on interlaced cross-attention attention musical feature fusion. The specific model architecture is illustrated in Figure. 1.

2.1 Multimodal representations module

The main idea of the multimodal representations module is to utilize the features of accompaniment audio to reduce bias in genre fusion decision results. The accompaniment of music (i.e., non-vocal parts) is primarily composed of instrumental and rhythmic elements, typically containing more musical features such as rhythm, melody, and chord information. In contrast, the vocal parts generally focus more on the expression and emotional conveyance of the song. Using music accompaniment audio as one of the feature sources can more effectively capture musical characteristics.

Vocal separation technique In this work, we utilized the publicly available music separation tool Spleeter [9] to separate vocals (including humming and

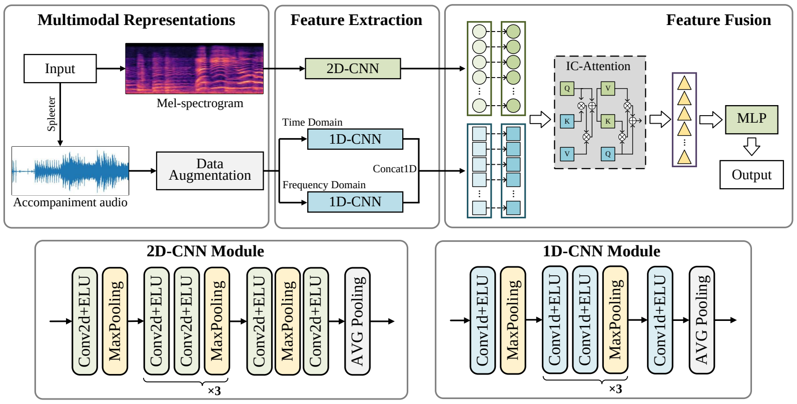

Fig. 1. The structure of ICFF-Net.

lyrics) and accompaniment (including melody and rhythm) from the audio data of the dataset. Spleeter is a deep learning-based music separation tool developed by Deezer, and its performance analysis will be introduced in Section 3.2.2. In the experiments, we employed a pre-trained 2-source model and fine-tuned it on our data.

Data augmentation module We augmented the accompaniment audio data using three methods to expand the dataset [14], thereby enhancing the model's robustness, generalization capability, and ability to handle complex data samples, enabling it to better adapt to various situations and variations. We considered these augmented accompaniment data as inputs for the 1D feature extraction task in our network.

Adding Noise We generate random numbers and convert them into uniform noise. This type of noise is a common random noise used to simulate randomness in signals or data in the real world. We randomly adjust the volume of the audio and mix the original accompaniment audio with the generated noise.

Time Shifting By moving the audio forward or backward by a certain time length, the temporal characteristics of the audio are changed without altering its pitch. This technique is achieved by randomly shifting the audio to the left or right. In this work, considering the presence of silent parts may lead to the loss of certain feature information, we set the audio to uniformly shift to the left by 1.5 seconds. The remaining 28.5 seconds of effective audio and 1.5 seconds of silent audio are cropped into 9 segments of 3 seconds each and added to the dataset.

Time Stretching We use the time-stretching technique from the audio processing library SoundTouch, which adjusts the duration of audio frames without altering the pitch, generating audio with a similar texture. The audio signal is first decomposed into small time windows, or frames, and analyzed using Fast

Fourier Transform (FFT) to determine the main frequency components in the audio. The duration of each frame is then adjusted. We set the stretching ratio to 0.9 or 1.1, randomly stretching all 30 seconds of audio by a random ratio (0.9 or 1.1) to generate new audio.

2.2 Feature extraction module

We designed two convolutional neural networks for extracting features from spectrograms and audio, respectively, at the 2D and 1D scales, as illustrated in Figure. 2. This approach facilitates the comprehensive consideration of various features and their combinations, thereby enabling more accurate classification of fused and non-fused genres, meeting the required feature elements.

2D-feature extraction module To effectively extract features from spectrograms, we designed a convolutional neural network. After preprocessing, the RGB three-channel spectrogram $x_{2d} \in \mathrm{R}^{H \times W \times C}$ is input into the neural network, where H represents the height of the image, W represents the width of the image, and C represents the number of channels, with RGB channels representing the red, green, and blue channels, respectively.

Starting with initial downsampling composed of 3×3 convolutional layers and 2×2 max-pooling layers, eight convolutional layers and four pooling layers are utilized to gradually extract more abstract and complex local features. Except for the last convolutional layer with a 5×5 kernel, all convolutional layers have a 3×3 kernel, and the max-pooling layers have a 2×2 kernel. The Exponential Linear Unit (ELU) [8] is used as the activation function for all convolutional layers. ELU is known for its soft saturation property, enhancing tolerance to noise, thereby improving the stability and generalization capability of the model. Finally, feature integration is performed through global average pooling to obtain the image feature $u_2 \in \mathrm{R}^{H \times W \times C}$.

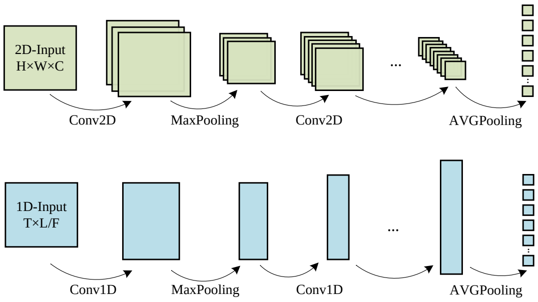

Fig. 2. The process of 2D and 1D feature extraction.

1D-feature extraction module The time-domain representation of audio captures the amplitude variations and waveform shapes of the audio signal over time, while the frequency-domain representation reveals the energy distribution or amplitude spectrum of the signal at different frequencies. This enables a more comprehensive understanding of the temporal and frequency characteristics of the audio signal, thereby more accurately capturing musical features. To fully consider the musical characteristics of audio, we perform feature extraction on both the time and frequency domains.

The original audio signal is used as the time-domain input $x_{t1} \in \mathrm{R}^{T \times L}$, where T represents the number of time frames, and L represents the number of samples or the length of the signal for each time frame. Simultaneously, the original audio signal is decomposed into localized segments in both time and frequency through short-time Fourier transform (STFT), resulting in the frequency-domain input $x_{f1} \in \mathrm{R}^{T \times F}$, where T is the number of time frames and F is the number of frequency components for each time frame. The expression for the STFT process is given by:

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n-m] \cdot e^{-j\omega n} \tag{1}$$

where m represents the starting position of the window, ω represents frequency, $x[n]$ is the original time-domain signal, $w[n-m]$ is the window function, j is the imaginary unit, and $e^{-j\omega n}$ is the complex exponential of frequency ω.

Similar to the 2D feature extraction, we constructed a 1D convolutional neural network (1D-CNN) for extracting 1D features, obtaining time-domain features $u_{t1} \in \mathrm{R}^{T \times L}$ and frequency-domain features $u_{f1} \in \mathrm{R}^{T \times F}$. Subsequently, we concatenated u_{t1} and u_{f1} along the third dimension (channel dimension) through 1D-concatenation, thus unifying the dimensions with u_2. The concatenated feature $u_1 \in \mathrm{R}^{T \times P \times 2}$, where $P = max(L, F)$, and $min(L, F)$ is estimated by linear interpolation to fill in the remaining time steps. This method preserves the smoothness of the data and retains the original trends of the data as much as possible.

2.3 Feature fusion and classification strategy

Interlaced cross-attention module To address the bias in fused genre classification, it's essential to carefully consider the interaction between features and the allocation of weights. Accurate classification of fused genres requires a high level of refinement in feature processing, and feature fusion can integrate different feature sources and allocate weights to adjust the requirements of different genre categories for features, achieving more diverse weight distributions to better adapt to the diversity and complexity of music [2]. Cross-attention is a feature fusion method based on attention mechanisms, allowing the model to better understand the relationship between different features and facilitate information interaction between features. Single cross-attention achieves feature fusion by guiding the attention or weight allocation of one feature with another feature, which typically focuses on one feature. Therefore, we constructed an interlaced

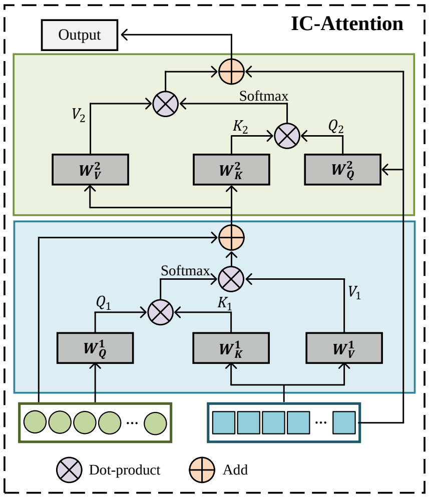

Fig. 3. The structure of IC-Attention.

cross-attention module with a two-layer structure, achieving comprehensive feature interaction and weight allocation through two embedding cross-attentions with different sources.

We propose IC-Attention to comprehensively consider the impact of both 2D and 1D audio features on classification results. IC-Attention consists of two consecutive cross-attention decoder layers, as illustrated in Figure. 3. In the first layer of this module, 1D features are used to enrich 2D features, while in the second layer, 2D features enrich 1D features, facilitating effective information interaction and fusion between features.

In Section 2.2, we obtained 2D features $u_2 \in \mathrm{R}^{H \times W \times C}$ and 1D features $u_1 \in \mathrm{R}^{T \times P \times 2}$ as inputs to the IC-Attention module. In the first layer of IC-Attention (highlighted in blue in the diagram), we use u_2 and u_1 as embedding representations, and then obtain query Q_1, key K_1, and value V_1 through different linear transformations as follows:

$$Q_1 = u_2 \times W_{Q1} \tag{2}$$

$$K_1 = u_1 \times W_{K1} \tag{3}$$

$$V_1 = u_1 \times W_{V1} \tag{4}$$

where W_{Q1}, W_{K1}, and W_{V1} are learnable weight matrices.

Next, we compute the attention scores α_1 by taking the dot product between Q_1 and K_1 as follows:

$$\alpha_1 = Softmax\left(\frac{Q_1\left(K_1\right)^T}{\sqrt{d_k}}\right) \tag{5}$$

where d_k is the dimensionality of the key vectors.

After obtaining the attention scores α_1, we perform matrix multiplication between α_1 and the value matrix V_1, and then add the result to u_2 to generate the 2D features enriched by the 1D features u_2' as follows:

$$u_2' = \alpha_1 V_1 + u_2 \tag{6}$$

In the second layer of IC-Attention (highlighted in green in the diagram), we adopt a similar approach. We use u_1 and u_2' as embedding representations and obtain query Q_2, key K_2, and value V_2 through linear transformations as follows:

$$Q_2 = u_1 \times W_{Q2} \tag{7}$$

$$K_2 = u_2' \times W_{K2} \tag{8}$$

$$V_2 = u_2' \times W_{V2} \tag{9}$$

where W_{Q2}, W_{K2}, and W_{V2} are learnable weight matrices. We then compute the attention scores α_2 by taking the dot product between Q_2 and K_2.

After obtaining the attention scores, we perform matrix multiplication between α_2 and the value matrix V_2 and then add the result to u_1 to generate the 1D features enriched by the 2D features u' as the output as follows:

$$u' = \alpha_2 V_2 + u_1 \tag{10}$$

In this way, the output u' achieves an implicit fusion of 2D and 1D features, considering the interrelation of features with another dimensional feature in an interlaced manner, achieving weight allocation and fusion of features.

Classification strategy We trained a single-hidden-layer MLP network for the final classification task. The hidden layer consists of 1000 nodes, with a dropout rate of 0.3, and employs the ELU activation function.

Table 1. The genre labels and the actual number of songs used in the FusionMusic.

GTZAN-f		ISMIR2004-f	
Genre	Track number	Genre	Track number
Metal	100	Classical	634
Rock	100	Electronic	221
Disco	100	Metal-Punk	90
Hip-hop	100	Rock-Pop	203
Country	100	World	244

3 Experiments

3.1 Datasets

In this work, we evaluate the model performance using the GTZAN and ISMIR2004 datasets for general music genre classification, while the FusionMusic dataset is specifically designed to assess the model's ability to differentiate fused genres. Details of the three datasets are provided below:

GTZAN GTZAN is a publicly available music genre classification dataset created by Tzanetakis [20]. The dataset consists of 1000 30-second music clips evenly distributed across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. In this work, we adopt a random 8:2 split to divide the dataset into training and testing sets.

ISMIR2004 The ISMIR2004 dataset [4] comprises 1458 music tracks categorized into 6 genres: classical, electronic, jazz blues, metal punk, rock pop, and world. The dataset is divided into 729 tracks for training and the remaining for testing. We removed tracks shorter than 30 seconds and extracted 30-second segments from each track for experimentation.

FusionMusic We constructed a dataset specifically for evaluating the model's ability to distinguish fused genres. It comprises subsets from the two aforementioned public datasets, GTZAN-f and ISMIR2004-f, each containing only fused genres. This dataset is not used for comparison with other models but solely for analyzing the classification bias of ICFF-Net towards fused genres. Details of the dataset, including genre labels and the actual number of songs used, are provided in Table 1.

3.2 Experiment settings

Mel-spectrogram In this study, we utilized the Librosa library to generate the mel-spectrogram of audio. First, the audio files are read in a mono channel with a sampling rate of 44.1kHz. Then, we divide it into short time windows and create a matrix of 128 mel-filters. Next, we perform STFT on the audio signal, convolve the resulting spectrogram with the mel-filter bank to obtain energy values and generate the mel-spectrogram. Finally, we plot a 2560×256 spectrogram image based on the energy spectral density in time and frequency and crop it into ten 256×256 segments as inputs for the network.

Vocal separation tool The choice of vocal separation tool directly affects the quality of the accompaniment audio, thereby impacting the subsequent 1D feature extraction input. Its performance is crucial for the entire model. In this work, we selected the publicly available music separation tool Spleeter. To ensure the accuracy and reliability of the experiment, we downloaded 50 songs and their accompaniment versions from NetEase Cloud Music. By using Spleeter, we obtained satisfactory separation results as shown in Figure. 4, validating its reliability and efficiency in audio processing. Therefore, we chose Spleeter as the vocal separation tool and fine-tuned it using a pre-trained 2-source model in the experiment.

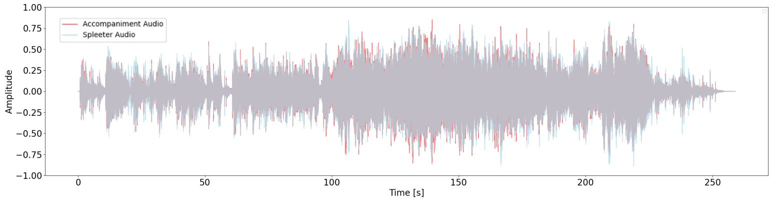

Fig. 4. The separation results of a Folk music. The red waveform represents the original accompaniment audio, the blue waveform represents the accompaniment audio separated by Spleeter, and overlapping parts are shown in gray.

Evaluation metrics In this study, accuracy, test loss, and F1-score were utilized as evaluation metrics to assess the classification performance of the proposed methods. Additionally, confusion matrices were used to intuitively illustrate the impact of classification bias on our model. Experiments in section 4.4 were conducted to analyze the classification impact of specific fused category groups that are easily confused.

Comparative experimental groups Our model was compared with ten MGC models to evaluate the classification performance. The models compared with ICFF-Net were as follows.

MCLNN [16]. The Masked Conditional Neural Network (MCLNN) aims to leverage the spatiotemporal nature of sound representation. It constrains inference based on prior and posterior time slices at specific time constraints. MCLNN employs controlled system sparsity and embeds behavior similar to a filter bank in the network.

FusionNet [18]. This network combines Convolutional Neural Networks (CNN) with NetVLAD and self-attention to capture local information across layers and learn their long-term dependencies. A meta-classifier learns aggregated high-level features from different local feature encoding networks for final classification.

MSNet [6]. Combines 1D SincNet with 2D ResNet, enhancing feature learning with Spatial Pyramid Pooling.

ASANet [3]. A framework that combines auditory image features with traditional acoustic and spectrographic features. The process involves extracting auditory image features based on the auditory image model, simulating the auditory system of the human ear.

S3T [23]. A self-supervised pre-training method for music classification using Swin Transformer, aiming to learn meaningful music representations from large volumes of easily accessible unlabeled music data. S3T offers a music data augmentation pipeline and two specifically designed preprocessors.

ACMNet [12]. Introduces a Soft Maximum Margin (AMCM-Softmax) method combining angular and cosine margins to enhance intra-class compactness and inter-class differences simultaneously. It eliminates radial variations by normalizing weight and feature vectors. Then, it introduces angular and cosine boundary parameters to maximize the decision boundary by applying angular and cosine boundary constraints.

MSDNN [21]. The model utilizes the HPSS (Harmonic-Percussive Source Separation) algorithm to decompose the spectrogram of the original music signal into two components: the harmonic feature component representing the tonal aspects and the percussive feature component representing the rhythmic aspects.

WSANet [10]. A hybrid deep learning model is employed for the analysis and classification of various music genre files. This hybrid model primarily combines multimodal models with transfer learning techniques for classification purposes.

ETNet [19]. The model proposes five innovative music classification methods: WVG-ELNSC, SDA classifier, RA-TSM transfer learning, TSVM algorithm, and BiLSTM and GCN deep learning classifiers.

RGCNN [22]. The model constructs an MSD algorithm based on CNNs and utilizes Harmonic-Percussive Source Separation (HPSS) to separate the spectrogram of the original music signal into two components: time-domain harmonic features and frequency-domain percussive features.

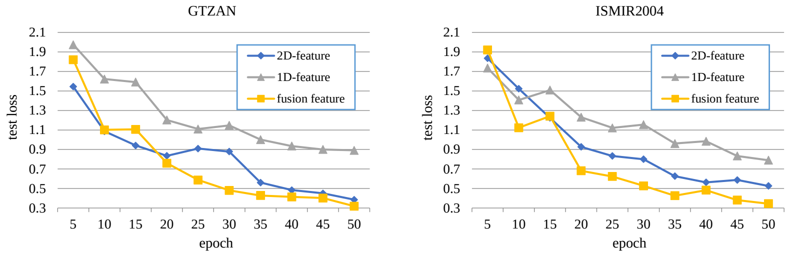

Fig. 5. Comparison of test loss with 2D-feature, 1D-feature, and fused feature in the GTZAN dataset and ISMIR2004 dataset. The horizontal axis represents training epochs, while the vertical axis represents test loss.

4 Experimental results

4.1 Comparative analysis: Original feature vs. Fused features

In our comparative experiments on the GTZAN and ISMIR2004 datasets, we employed 2D feature u_2, 1D feature u_1, and their fused feature u' as inputs to an MLP layer for classification. Notably, in the experiments with fused features, 1D audio data underwent no data augmentation.

As shown in Figure. 5, the results indicate that using fused features yields superior performance compared to using either 2D or 1D features alone, as evident from the test loss. Using 1D features for classification yielded significantly

Table 2. Comparison of the performance of different state-of-the-art methods on the GTZAN and ISMIR2004 datasets.

	GTZAN			ISMIR2004		
Models	Accuracy	Test-loss	F1-score	Accuracy	Test-loss	F1-score
MCLNN[16]	0.9025	0.3152	0.8936	0.8604	0.4216	0.8651
FusionNet[18]	0.9114	0.2964	0.9139	0.9246	0.2456	0.9210
MSNet[6]	0.9149	0.2963	0.9205	0.9191	0.2687	0.9206
ASANet[3]	0.9180	0.2899	0.9126	0.8290	0.4967	0.8122
S3T[23]	0.8110	0.5027	0.8090	0.8021	0.5521	0.8201
ACMNet[12]	0.7450	0.7592	0.7562	0.7329	0.8362	0.7263
MSDNN[21]	0.8496	0.4547	0.9525	0.8301	0.4862	0.8467
WSANet[10]	0.8100	0.5592	0.7295	0.7953	0.6228	0.7058
ETNet[19]	0.9301	0.2623	0.9012	0.9028	0.2684	0.9191
RGCNN[22]	0.9309	0.2548	0.9180	0.8571	0.4342	0.8615
ICFF-Net(ours)	0.9312	0.2486	0.9324	0.9194	0.2708	0.9266

worse results compared to using 2D features alone, indicating that while 1D features may contain temporal information and relevant musical features, they are unable to capture more complex features. When used as auxiliary features providing temporal information, they can offer a more comprehensive feature representation, leading to improved classification performance. It also demonstrates that fused features have better classification performance compared to single features.

4.2 Comparison with the state-of-the-art

As shown in Table 2, our approach achieves the highest accuracy on the GTZAN dataset, reaching 93.12%, while FusionNet proposed demonstrates the best classification performance on ISMIR2004. FusionNet integrates eight diverse features, yielding optimal results. However, when considering fusion with only two types of features, it reaches an accuracy of 87.38%, which is inferior to our method. Furthermore, our model achieved the highest F1-score on two datasets, reaching 93.24% and 92.66%.

ICFF-Net excels on the GTZAN dataset, benefitting from its diverse fusion genres. Conversely, ISMIR2004 merges similar genres, mitigating fusion issues. Additionally, it includes instrumental tracks, where our augmentation strategy offers marginal improvements. Overall, our model enhances classification accuracy, addressing fusion genre classification bias, discussed in Section 4.4.

4.3 Ablation study

As shown in Table 3, combining the three audio augmentation strategies reveals that the addition of noise is less effective individually, while time stretching

Table 3. Classification results of combined audio augmentation methods: o - original audio without data augmentation, a - noise addition, b - time shifting, and c - time stretching.

Method		GTZAN			ISMIR2004		
		Accuracy	Test-loss	F1-score	Accuracy	Test-loss	F1-score
o	cross-attention	0.8606	0.4398	0.8569	0.8703	0.3935	0.8804
	IC-Attention	0.8746	0.3958	0.8719	0.8826	0.3859	0.8912
a	cross-attention	0.8715	0.3952	0.8756	0.8720	0.3697	0.8865
	IC-Attention	0.8769	0.3903	0.8805	0.8727	0.3891	0.8806
b	cross-attention	0.8947	0.3493	0.8723	0.8774	0.3693	0.8902
	IC-Attention	0.9065	0.3156	0.8939	0.8869	0.3491	0.8706
c	cross-attention	0.9124	0.2912	0.9003	0.8803	0.3515	0.9086
	IC-Attention	0.9146	0.2880	0.9025	0.9135	0.2758	0.9005
ab	cross-attention	0.8798	0.3616	0.8625	0.8771	0.3567	0.8765
	IC-Attention	0.8864	0.3594	0.8608	0.8710	0.3496	0.8736
ac	cross-attention	0.8722	0.3952	0.8836	0.8759	0.3737	0.8869
	IC-Attention	0.8865	0.3798	0.8905	0.8895	0.3653	0.9036
bc	cross-attention	0.9171	0.3008	0.9039	0.8875	0.3419	0.8969
	IC-Attention	0.9274	0.2924	0.9136	0.9059	0.3025	0.9065
abc	cross-attention	0.8903	0.3309	0.9068	0.8948	0.3396	0.8841
	IC-Attention(ours)	**0.9312**	**0.2486**	**0.9324**	**0.9194**	**0.2708**	**0.9266**

performs better. This may be because adding noise results in audio similar to the original, making 1D feature extraction less effective. However, combining all three techniques increases the original data nearly fourfold and yields the best classification results.

For the GTZAN dataset, in most studies [3], the accuracy of Rock is consistently the lowest. It is easily confused with other genres, consistent with our understanding of fusion genres. As shown in the confusion matrix in Table 4, our model significantly improves the discriminative accuracy of fusion genres while also enhancing overall classification accuracy. This demonstrates the effectiveness of combining 2D and 1D features and considering the correlation between different features in addressing fusion genre classification bias.

4.4 Analysis of classification bias on fusion genres

We constructed the FusionMusic dataset to evaluate the model's ability to differentiate fusion genres. The dataset includes GTZAN-f and ISMIR2004-f subsets, containing only fusion genres. We utilized 2D feature u_2, 1D feature u_1, and their fused feature u' as inputs to an MLP layer for classification, with u' representing the ICFF-Net scenario.

As depicted in Figure. 6, each fusion genre exhibits specific misclassification tendencies. In GTZAN-f, Rock is prone to confusion with several other genres; when using individual features for classification, Rock tends to be misclassified as

Table 4. Confusion Matrices (%) of ICFF-Net in GTZAN dataset.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0)Blues	**95**	0	2	1	0	0	0	0	0	3
(1)Classical	0	**96**	0	0	4	0	0	0	0	0
(2)Country	0	2	**91**	2	0	0	0	1	1	3
(3)Disco	0	0	1	**97**	0	0	0	2	0	0
(4)Hiphop	0	0	1	0	**94**	0	0	2	1	3
(5)Jazz	2	4	3	1	0	**90**	0	0	0	0
(6)Metal	0	0	0	0	5	0	**89**	0	0	6
(7)Pop	0	0	0	1	1	0	0	**93**	3	0
(8)Reggae	2	0	0	0	1	1	0	2	**94**	0
(9)Rock	1	0	2	2	0	0	2	0	1	**92**

Hiphop and Disco, while Metal and Country are more often misjudged as Rock. Musically, Rock shares rhythmic emphasis with Hiphop and Disco, while Metal songs often structurally resemble classic Rock tunes. Additionally, both Country and Rock extensively feature guitars, especially in subgenres like Country Rock and Southern Rock, blurring the line between them.

These similarities in instrument use, rhythmic structure, emotional expression, and performance style make them difficult to classify accurately. However, employing ICFF-Net significantly improves this situation. In ISMIR2004-f, Pop-Rock and World genres are more prone to misjudgment, once again highlighting the impact of Rock as a typical fusion genre in classification tasks. Similarly, using ICFF-Net reduces inter-genre misjudgments in this dataset.

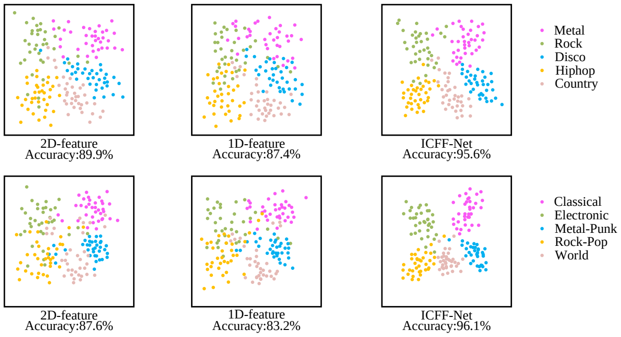

Fig. 6. In the FusionMusic dataset, we visualize the classification results of ICFF-Net to demonstrate the effect of model bias in classifying fusion genres. The upper figure shows the classification results on GTZAN-f, while the lower figure shows the classification results on ISMIR2004-f.

5 Conclusion

In this paper, we introduced the ICFF-Net, an interlaced cross-attention fusion feature network designed to address bias issues in fusion music classification. The ICFF-Net extracts features from both 1D and 2D sources, combining temporal and frequency features of 1D with 2D features. The interlaced cross-attention module, featuring a dual-layer structure, effectively fuses these feature types, promoting comprehensive feature interaction and weight allocation. The final classification is determined using MLP. Three data augmentation methods were employed to expand the accompaniment audio dataset, leading to the creation of the FusionMusic dataset for experimental analysis of fusion genres. Through comparisons with state-of-the-art methods and discussions on classification bias, our model demonstrated significant advantages in improving classification accuracy and addressing bias among fusion genres.

In this study, we utilized 2D-CNN and 1D-CNN networks as feature extraction modules, proving sufficient for extracting necessary features for classification. Future work will explore feature extraction modules with more complex structures, incorporating deeper networks such as ResNet, DenseNet, ConvNeXt, to enrich feature inputs from different sources. Additionally, diverse feature fusion methods will be investigated to further mitigate bias in fusion genre classification.

References

1. Allamy, S., Koerich, A.L.: 1d CNN architectures for music genre classification. In: IEEE Symposium Series on Computational Intelligence, SSCI 2021, Orlando, FL, USA, December 5-7, 2021. pp. 1–7 (2021)
2. Bishop, C.M.: Pattern recognition and machine learning, 5th Edition. Springer, London (2007)
3. Cai, X., Zhang, H.: Music genre classification based on auditory image, spectral and acoustic features. Multim. Syst. **28**(3), 779–791 (2022)
4. Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., Wack, N.: Ismir 2004 audio description contest (2006)
5. Chaki, J.: Pattern analysis based acoustic signal processing: a survey of the state-of-art. Int. J. Speech Technol. **24**(4), 913–955 (2021)
6. Chang, P., Chen, Y., Lee, C.: Ms-sincresnet: Joint learning of 1d and 2d kernels using multi-scale sincnet and resnet for music genre classification. In: ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021. pp. 29–36 (2021)
7. Cheng, X., Shi, F., Zhang, Y., Li, H., Liu, X., Chen, S.: Frame: Feature rectification for class imbalance learning. IEEE Transactions on Knowledge and Data Engineering (2024)
8. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)

9. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. J. Open Source Softw. **5**(56), 2154 (2020)
10. Jena, K.K., Bhoi, S.K., Mohapatra, S., Bakshi, S.: A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis **35**, 11223–11248 (2023)
11. Kim, J., Won, M., Serra, X., Liem, C.C.S.: Transfer learning of artist group factors to musical genre classification. In: Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018. pp. 1929–1934 (2018)
12. Li, J., Han, L., Wang, Y., Yuan, B., Yuan, X., Yang, Y., Yan, H.: Combined angular margin and cosine margin softmax loss for music classification based on spectrograms. Neural Comput. Appl. **34**(13), 10337–10353 (2022)
13. Li, S., Cheng, X., Shi, F., Zhang, H., Dai, H., Zhang, H., Chen, S.: A novel robustness-enhancing adversarial defense approach to ai-powered sea state estimation for autonomous marine vessels. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2024)
14. de Lima Aguiar, R., Costa, Y.M.G., Jr., C.N.S.: Exploring data augmentation to improve music genre classification with convnets pp. 1–8 (2018)
15. Lu, L., Zhang, H., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. Multim. Syst. **8**(6), 482–492 (2003)
16. Medhat, F., Chesmore, D., Robinson, J.: Masked conditional neural networks for sound classification. Appl. Soft Comput. **90**, 106073 (2020)
17. Meng, S., Hao, Q., Xiao, Y., Zheng, W.: Musicnext: Addressing category bias in fused music using musical features and genre-sensitive adjustment layer. Intelligent Data Analysis (Pre-press), 1–15 (2023)
18. Ng, W.W.Y., Zeng, W., Wang, T.: Multi-level local feature coding fusion for music genre recognition. IEEE Access pp. 152713–152727 (2020)
19. Prabhakar, S.K., Lee, S.W.: Holistic approaches to music genre classification using efficient transfer and deep learning techniques **211**, 118636–118636 (2023)
20. Tzanetakis, G., Cook, P.R.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
21. Wang, X.: Music similarity detection guided by deep learning model. **2023**, 1263620–10 (2023)
22. Xie, C., Song, H., Zhu, H., Mi, K., Li, Z., Zhang, Y., Cheng, J., Zhou, H., Li, R., Cai, H.: Music genre classification based on res-gated cnn and attention mechanism pp. 13527–13542 (2024)
23. Zhao, H., Zhang, C., Zhu, B., Ma, Z., Zhang, K.: S3T: self-supervised pre-training with swin transformer for music classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. pp. 606–610 (2022)
24. Zhao, Z.: Music classification model based on feature fusion. In: 18th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2019, Beijing, China, June 17-19, 2019. pp. 426–429 (2019)