International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1159-1169

https://doi.org/10.70517/ijhsa464108

Research on spatio-temporal feature extraction and algorithm optimization in music composition under traditional culture

Yina Jia[1,*]
[1] College of Music, Changchun University, Changchun, Jilin, 130022, China
Corresponding authors: (e-mail: yinajia2024@163.com).

Abstract Music composition under traditional culture not only involves complex artistic expression, but also needs to be combined with modern computational methods to improve the efficiency and quality of composition. This study proposes a music composition model based on the Improved Multitrack Sequence Generation Adversarial Network (RFGAN), which aims to improve the quality and coherence of the generated music. The model optimizes the music generation process by introducing a loop-structured generator and timing model, combined with a discriminative feedback mechanism inside and outside the tracks. Comparative experiments were conducted to evaluate the model's note prediction using Top1, Top2 and Top3 accuracies, and the results showed that RFGAN achieved the highest 88.79% Top1 accuracy in note prediction. To further validate the effectiveness of the model, the study also used a twelve-mean rhythm comparison, and the results showed that the generated note distributions were similar to the real music data, indicating that the model was able to capture the regularity of the music. In addition, the music generated by the model also outperforms the traditional GAN and BiGRU models in terms of harmony, rhythm, and overall effect, verifying its advantages in music composition.

Index Terms Music Composition, Generative Adversarial Networks, Note Prediction, Track Generation, Harmonization, Temporal Modeling

I. Introduction

Traditional culture is often the cultural foundation and cultural symbol of a nation, with the development of modern science and technology, the improvement of productivity level, the penetration of foreign culture, a considerable portion of the traditional culture of the counteracting effect is more and more insignificant, more and more do not adapt to the contemporary socio-economic development, is in an alarming rate of extinction [1]-[4]. While music, as an art form, is inseparable from culture in historical development, applying traditional culture to music creation not only helps to inherit and promote traditional culture, but also increases the cultural connotation and artistic value of music [5]-[7]. It helps to cultivate the public's awareness and understanding of traditional culture and revitalize its vitality and life in popular music [8].

In music composition, traditional culture is utilized in various ways [9]. Among them, the most common is the tracing and borrowing of traditional culture and musical instruments [10]. For example, using the characteristic instruments such as guqin and silk and bamboo instruments in folk music to create music, fusing modern electronic elements with traditional instruments, it can make the music have new and vivid elements with a stronger cultural charm [11]-[13]. In addition, artists also search for inspiration from traditional culture and integrate some of these myths and legends or traditional stories into music [14], [15]. At present, many creators pay more attention to listenability and novelty in music creation, while neglecting the value and significance of traditional culture [16]. This situation makes many musical works lack the depth and connotation of traditional culture, and it is more difficult to be remembered and inherited by future generations [17]. Therefore, in future music creation, the combination of traditional culture and music creation should be emphasized, and various traditional cultural elements should be integrated into music creation to create music works with ancient style and traditional cultural characteristics [18]-[20]. This can not only make the music more connotation and depth, but also make a positive contribution to the inheritance and promotion of traditional culture for the country [21], [22].

This study focuses on the optimization problem of multi-track music composition based on generative adversarial networks (GANs). Although traditional GAN models are capable of generating simpler note sequences, they perform poorly in dealing with complex musical structures, especially in terms of note coherence and harmonization among multi-tracks. Therefore, this study proposes an improved RFGAN-based model, which not only enhances the articulation between notes, but also better simulates the rhythmic and harmonic aspects of music composition by introducing a loop structure and a timing model. In addition, the model employs a feedback mechanism inside and outside the track, which improves the quality of the generated music composition. Specifically, the research methodology includes preprocessing of the dataset, feature extraction, model training and evaluation of the generated results. By using the main melody tracks of 800 popular songs as training data, the improvement of note prediction accuracy was emphasized, and comparative experiments were employed to verify the effectiveness of the new model. The results show that RFGAN performs well in note generation, track coordination, and music style reproduction, providing strong technical support for future intelligent music creation.

II. Techniques related to music composition under traditional culture

In the field of sequence music creation, the main difficulty faced is that the length of the generated sequence is constrained by the maximum upper limit of the model, which makes it difficult to generate music sequences that are longer and at the same time more qualitative, and many models begin to generate music that is disorganized and irregular as the length of the generated sequences increases in the automatic creation of the generation. At the same time, in the field of sequence generation, it is also accompanied by the problem of slow inference.

II. A.Theoretical knowledge of music composition

II. A. 1) Theoretical knowledge of music

Pitch is the position of an individual sound in music across the entire range of sound frequencies. The pitch of sounds depends on the frequency of vibration of the sound waves that produce them. Higher frequencies are considered higher pitches, while lower frequencies are considered lower pitches. Pitch is one of the most important attributes of music that can act directly on the human ear's perceptual organs and is a visual representation of the small grain size of music. In computers, in addition to recording music in wav, there is another way to quantize and save pitch using a discrete quantization form, which is divided into a pitch range of [0,126] to represent pitch [23].

Intervals refer to the interrelationship of two levels in pitch, that is, in terms of the distance between two tones in pitch, and the name of its unit is called degree. Intervals are divided into concordant and discordant intervals, which are actually distinguished by vibrational ratios. There are three types of harmonic intervals, namely, perfect harmonic intervals, perfect harmonic intervals, and imperfect harmonic intervals. Extremely fully concordant intervals are the pure 1st degree and almost fully united pure octaves, fully concordant intervals are the pure 5th and pure 4th degrees which are quite integrated in sound, and incompletely concordant intervals are the large and small 3rd and large and small 6th degrees. The dissonant intervals are harsh and discordant, the major and minor second, the major and minor seventh and all increasing and decreasing intervals are dissonant intervals. The intervals of music occupy an important position in musical creation, and a beautiful piece of music must maintain good intervals in the intervals. In the music of the light and heavy, before and after echo, emotional expression and other creative ideas, can be set through the appropriate interval to realize.

A melody, also called a tune, is an organized sequence of musical notes. It is a line of tones composed according to a certain relationship of height, length, strength and weakness. It is the most important means of shaping the image of music and is the soul of music. Compared with other elements, melody has the most prominent position in music, and naturally it is the most important means of expression. It can be said that the first thing anyone remembers after listening to a piece of work is the melody, which can be hummed by us, while other elements such as mere register, intensity, rhythm, leaving the melody, simply can not exist independently, so it has been said that the melody is the soul of music.

II. A. 2) Music notation

In current research such as music generation through deep learning, music usually exists in both Symbolic and Audio forms.

In Symbolic form, the music is composed of a series of notes, each of which holds information such as pitch, note onset time, termination event, duration, intensity, instrument type, etc. This form of music is usually saved in MIDI format or Piano Roll. Music in this form is usually saved in MIDI format or Piano Roll.

(1) Musical Instrument Digital Interface (MIDI) is a widely used standard format for music, and the MIDI format is very small; MIDI uses notes as the basic unit, and MIDI events include note on and note off, with the time interval between them being the duration of a note, and each note also has a pitch, ranging from 0 to 126. In addition to this, MIDI also stores information such as the intensity of the note [24].

(2) Piano Roll, a two-dimensional table with a time frame on the horizontal axis and pitch on the vertical axis, is simpler and holds less information. It is also one of the commonly used music representations.

Music saved in Symbolic form occupies very little space while retaining the main features of the music, which can significantly reduce the complexity of the data and the amount of computation in tasks such as music generation. Music generation tasks that use Symbolic as a data representation generally also generate music in Symbolic form, which is currently used when generating music through Generative Adversarial Networks and Recurrent Neural Networks.

The Audio form encodes all the information needed to realize a piece of music, including the stylistic differences of the musicians, and is very complete, but does not explicitly represent the note-related information that is emphasized in the Symbolic, such as the start time and end event of the note. This form of music saves waveforms or Fourier transformed spectrograms, etc. Audio saves a lot of information, occupies a large amount of memory, and is computationally intensive. However, the use of Audio can generate music directly, reducing the intermediate process, which makes the data processing much simpler, and at the same time can generate a variety of music, even the style of the specified musician's music, which is difficult to do with Symbolic. Currently, the models that use Audio form are usually WaveNet and related improved versions such as WaveGAN and other methods.

II. B.Multi-track sequence generation adversarial network

II. B. 1) Generating Adversarial Network Principles

The basic idea behind neural networks, a common machine learning model, is to use high-dimensional nonlinear functions to estimate or approximate modeling of data. Deep learning, on the other hand, is an algorithm that uses neural networks as an infrastructure to learn the features implicit in the data.

Generative Adversarial Networks (GANs) are mainly inspired by the idea of zero-sum games in game theory, and their basic structure is shown in Fig. 1, which contains two parts, the generator (G) and the discriminator (D) [25]. This model is generally realized through neural networks, which can be implemented using deep neural networks with deep structure as an architecture, in addition to any form of differentiable system capable of mapping data in a space to other spaces.

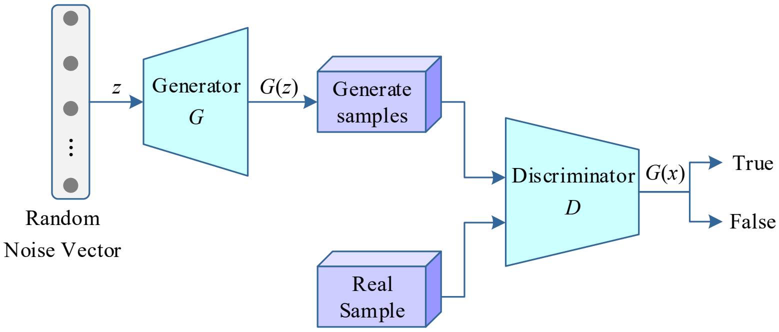


Figure 1: The basic structure of GAN

Where the generator samples randomly from the latent space, i.e., with a random noise vector as input, and its goal is to capture and fit the distribution of the real samples, generating data samples G(z) that are as similar as possible to the real samples, in order to deceive the discriminator as much as possible. The discriminator, which is usually a binary classifier, takes as input either the real samples or the data samples generated by the generator, and its goal is to distinguish the generated samples from the real samples as accurately as possible. The two models are constantly fighting against each other and adjusting their parameters, and their ultimate goal is to make the discriminator unable to accurately determine the authenticity of the generated samples, i.e., whether or not the generated samples are real samples.The loss function of the GAN during training is denoted as:

$$\begin{aligned}\min_{G}\max_{D} V\left(D,G\right) &= E_{x\sim p_{data}(x)}\left[\log D(x)\right] \\ &+E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]\end{aligned} \tag{1}$$

where $p_{data}(x)$ represents the data distribution of real samples and $p_z(z)$ represents the data distribution of random noise vectors.GAN learns by letting the generator and the discriminator play against each other, and the optimal tuning of the model parameters is a min-max optimization problem. The adversarial game between the generator and the discriminator and the tuning of the parameters terminate at a saddle point, which is the minimum for the generator, i.e., the minimization function $V(D,G)$, to minimize the difference between the generated samples and the real samples. And for the discriminator is the maximum value, i.e., maximizing the function $V(D,G)$, to improve the discriminator's ability to discriminate between the real samples and the generated samples as much as possible. In other words, the training and learning process of GAN is essentially a “binary minima maxima game” process, and its ultimate goal is to achieve Nash equilibrium. Then it can be assumed that the generator has learned the distribution of real samples, and can generate new data samples that are "real" enough.

II. B. 2) Multi-track sequential GAN modeling

The different tracks in each musical composition are independent and closely related, presenting their own music, yet unfolding interdependently over time. Thus, there are traditionally two main types of musical composition, depending on how the tracks are presented. One is that musicians playing different instruments create music by improvising on different instruments without music books, pre-rehearsal arrangements, or a musical conductor; this mode of composition can be called "Jamming". Another way of composition is to envision a composer who creates scores for different instruments in advance arrangements according to the harmonic structure of the music and the instrumentation of the instruments (i.e., an important part of the orchestration course teaching in composition majors), and then the composer follows his own compositional scores to direct different instruments to play the music together according to the pre-existing scores, which can be referred to as the compositional mode of "Composer".

In contrast to these two modes of composition, two types of discriminative feedback mechanisms have been proposed in Multitrack Sequential Generative Adversarial Networks (MuseGAN), one being a within-track discriminative feedback mechanism for the individual generation of each track, in the same way that each specific musical instrument is paired with a specialized music teacher. The other is an inter-track discriminative feedback mechanism, in which the discriminator plays the role of a composer or orchestra leader and attempts to correlatively evaluate the common performance of all musicians (i.e., tracks) in a musical composition.

In the Jamming model, multiple generators work independently and generate music belonging to their own tracks from different random vectors $z_{r,i} \in \{1,2,\ldots,m\}$, where m denotes the number of generators (or tracks).

In the Composer model, only one generator is needed to generate a multi-channel piano roll, where each channel in the piano roll represents a specific track.The Composer model requires only one input random vector, which can be thought of as the composer's idea of composing a song. The model requires only one discriminator, because the generator generates all tracks simultaneously from the random vector, and each track is interrelated.

By combining the ideas of both Jamming and Composer, one can further propose a hybrid (Hybrid) model that utilizes both intra- and inter-track discriminative feedback.The Hybrid model uses M generators to create M tracks of music, with each generator taking as input the inter-track random vector z and the intra-track random vector z_i. Also, the expectation is that the inter-track random vectors can coordinate the generation of different music (G_i), just like a composer, and thus only one discriminator is used to evaluate the M tracks together.

III. Music composition modeling based on improved multi-track sequential GANs

Today, with the rise of artificial intelligence, computerized music is developing rapidly. Music has two important aspects: composition and performance. Different genres of music convey different styles, and performers inject different rhythms and strengths into them when they play, thus generating rich expressiveness. The development of image style conversion has opened up the exploration of music style conversion, and the study of music style conversion can inspire music creators and assist music generation, which is an important issue in many fields of artificial intelligence.

III. A. Music Composition Dataset and Preprocessing

III. A. 1) Music composition dataset

The main research content of this paper is the generation of music composition for popular songs. Since there is no publicly available dataset of music for popular songs for the time being, this paper produces its own dataset by means of network collection. The dataset includes 800 songs, and the sheet music of each song is unified as a MIDI format score. The score contains two tracks, one is the main melody track and the other is the accompaniment track, of which the main melody track is selected as the model training dataset in this chapter. In order to ensure the uniformity of the note duration, the velocity of the scores are all set to 100 BPM.

III. A. 2) Pre-processing of data sets

In this paper, we design a music textualization representation for automatic music generation, which applies text processing methods to music, making it possible to transform melodies into one-dimensional music text signals without affecting the musical information. Music textualization makes the representation of the music dataset consistent with the inputs of most text generation models, which to some extent solves the current problem of different inputs for different models and the difficulty of model design in automatic music generation.

The preprocessing process for music data begins with converting the musical key and adjusting the tempo. Then the pitch and rhythm information of the notes are extracted to get the melody text. The specific steps are as follows:

(1) Transpose. The purpose is to convert all MIDI music to the same key and the same tempo, because the normalization of the data format helps the model to learn the characteristics of the music. All notes are converted to the key of C by tonal transposition (incrementing and decrementing all notes by intervals). After this, all melody track files had a dominant note of C (i.e., they were all music in the key of C), and the rest of the notes varied regularly around the dominant note. Then, the tempo of all MIDI music was set to 80, so that all MIDI music was played at the same tempo.

(2) Extracting note length information. Obtain the pitch length of the shortest note in the melodic score, and use it as the basic unit length to extract the pitch length information of all notes, which solves the problem of missing pitch length information in the conversion process.

(3) Extract pitch and rhythm information. The previously extracted unit length is used as the minimum length, and note pitch, note duration and pause duration are extracted at the same time.

Through the above pre-processing, all 800 popular songs were tuned to the key of C major. All the songs were then melodically textualized to obtain a series of musical melodic texts. The music texts were divided into 50 subsets to facilitate model training.

III. B. Improved GAN model construction for multi-track sequences

III. B. 1) Improvement of GAN with multi-track sequences

This paper proposes some improvements based on the multi-track sequence generative adversarial network model by analyzing the improvements that can be made to several widely used music generation models. By improving the structure of the network model into a recurrent generative adversarial network with a feature extractor, this model is able to improve the coherence between the generated music segments and the overall auditory effect of the music. At the same time, the connection between each data sample is increased, which makes the articulation between notes more natural and avoids the fragmentation problem. Specifically as follows:

(1) Generator with cyclic structure

In the traditional GAN generator structure, this paper has changed from a unidirectional structure to a cyclic structure, the unidirectional structure of the generator operation process is to send random noise vectors into the generator, the generator then generates the data samples, and then the data samples will only be input into the discriminator for judgment, and the data samples will no longer be used in the next round of training.

(2) Reward and punishment mechanism guided by music theory

In this paper, the cross-entropy loss function will be used as the loss function of the discriminator, which can well improve the quality of the generated music. At the same time, in order to make the generated music follow the rules of music theory, the basic music theory is modeled by means of mathematical modeling, and the importance of different contents in music theory varies in composing, which will be reflected in the reward and penalty value of the generated network, i.e., the more important contents bring higher reward and penalty values.

In order to make the generated music more in line with real music, the pitch range of the notes generated for different tracks should be set, i.e.:

$$R^{m_1}(S_{1:t},y_t)=\begin{cases}0.1, & y_t\in[y_{\min},y_{\max}]\\ -0.6, & y_t\notin[y_{\min},y_{\max}]\end{cases} \tag{2}$$

where $y_{\min}$ and $y_{\max}$ represent the lowest and the highest notes of the range, y_t represents the pitch value of the note at the t moment, and R represents the reward and penalty value.

In order to make the generated music more in line with the real playing situation, there should be a limit on the simultaneous sounding of notes in different tracks. To wit:

$$R^{m_2}(a_t)=\begin{cases}0.2, & a_t\le n\\ -0.5, & a_t>n\end{cases} \tag{3}$$

where a_t is the number of simultaneously voiced notes at t, n is the number of simultaneously voiced notes per track limit, and R is the reward and penalty value.

In order to make the generated music more close to the real music works, the number of rests used in each measure should be appropriately limited, here the number of rests is limited to not exceed 50% of the total number of notes in each measure. That is:

$$R^{m_3}(y)=\begin{cases}0.1, & y\le 0.5S_{1:t}\\ -0.3, & y>0.5S_{1:t}\end{cases} \tag{4}$$

where S denotes the number of notes in the measure, y denotes the number of rests, and R is the reward or penalty value.

(3) Timing model

The decoder is the mirror structure of the generation network model, and $E(y)$ is the output of the decoder, which is spliced with the random noise vector z into the Track-conditional generation timing model to generate the current round of data samples. Generation timing model to generate the current round of data samples $G(z,y)$. The expression of the temporal model is:

$$\begin{cases} G(z_i,y)=\{G^o(E(y),z)\}^T \\ y=G_{bar}(G_{temp}(z)^T),s=1 \\ y=G_{s-1}(z,y),\ s>1 \end{cases} \tag{5}$$

where s represents the number of rounds of training.

(4) Feature Extractor

A feature extractor is an algorithm that can extract useful information from the input music data by analyzing the time and frequency information of the music, as well as other features related to music composition. These features can be fed into the generative model to help the model better learn the laws of music composition. During the training process of the generative model, the music feature information extracted by the feature extractor is used to guide the learning of the generative model. The feature extractor can also compare the music results that have been generated with the original music data to evaluate the performance of the generative model and make adjustments and optimizations in the next round of training.

III. B. 2) RFGAN model network structure

Based on the improvement way about the multi-track sequence GAN model in Section 3.2.1, a new multi-track sequence music composition generation model, i.e., the Recurrent Feature Generation Adversarial Network model (RFGAN), is formed by combining the above improvement scheme, and its specific framework is shown in Fig. 2, which can be used for generating contextually relevant music phrases.

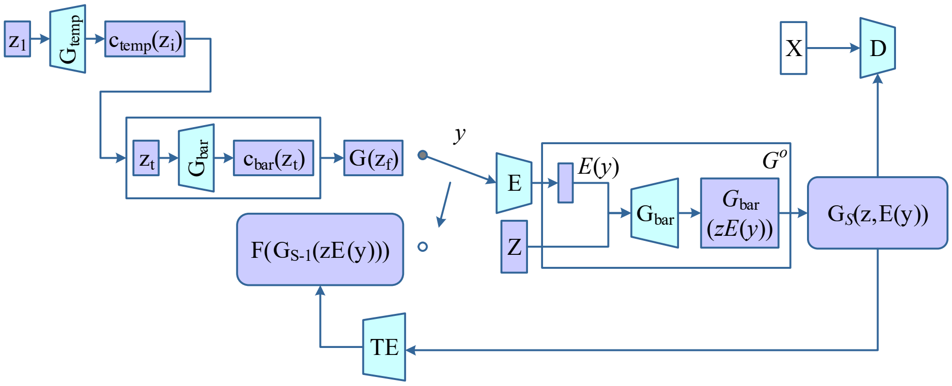


Figure 2: The RFGAN model summarizes the composition

The first part of the input is the main melody generation part as the a priori condition of the generative model. In the first round of training of the model, a monophonic music sample is generated from the T1 pattern, which is used as the main melody to be fed to the decoder E. In the subsequent rounds of training, the result samples of the previous round of training are used, $F(G_{s-1}(z,E(y)))$, and the feature information of the result samples is extracted by the feature extractor TE, and decoded by the decoder E, and fed into the convolutional layers of different generative models, respectively. $F(G_{s-1}(z,E(y)))$, the feature information is decoded by the decoder E and fed into the convolutional layers of different generative models (with the same shape of the input array matrices) to influence the generative models to produce music results. The result samples $G_s(z,E(y))$ from this round of training will not only be used as inputs to the discriminative model D along with the real training set x, but will also be used in the next round of generative model training. Same as MuseGAN model training strategy, according to every 6 updates of the discriminative model, the generative model is updated once. Training alternates in turns.

The input random noise Z in the T2 timing model is composed of four parts respectively four random noises within and between tracks. One part is the shared weights input to the five-track generation model, random noise z_t with time series and random noise z without time series, which are used to constrain the inter-track music generation.

One part is the random noise z_t and z with and without time series within the tracks, which is used to influence the music generation within the tracks. To wit:

$$y = G_{bar}(G_{temp}(z_i)^T), s = 1 \tag{6}$$

$$y = G_{s-1}(z, y), s > 1 \tag{7}$$

$$G(z, y) = \left\{G^o(E(f(G_{s-1}(z, y))), \tilde{z})\right\}^T \tag{8}$$

Eq. (6) is the expression for the total model generating model, and Eqs. (7) to (8) are the y expression functions for the case of different number of training rounds, respectively.

IV. Validation analysis of the validity of music composition modeling

With the improvement of living standards, people's needs in music are gradually increasing, such as game background music, movie and television episodes, theme songs, etc. Human compositions are far from meeting the market demand, so artificial intelligence compositions appear with it. Generally speaking, music can be divided into two types: mono-track and multi-track, and from the method of music generation, it is divided into symbolic music generation and audio music generation. How to effectively realize the feature extraction of music composition generation is an innovative direction to further improve the quality of music creation.

IV. A. Comparison of objective evaluation of musical composition

IV. A. 1) Comparison of creative effects

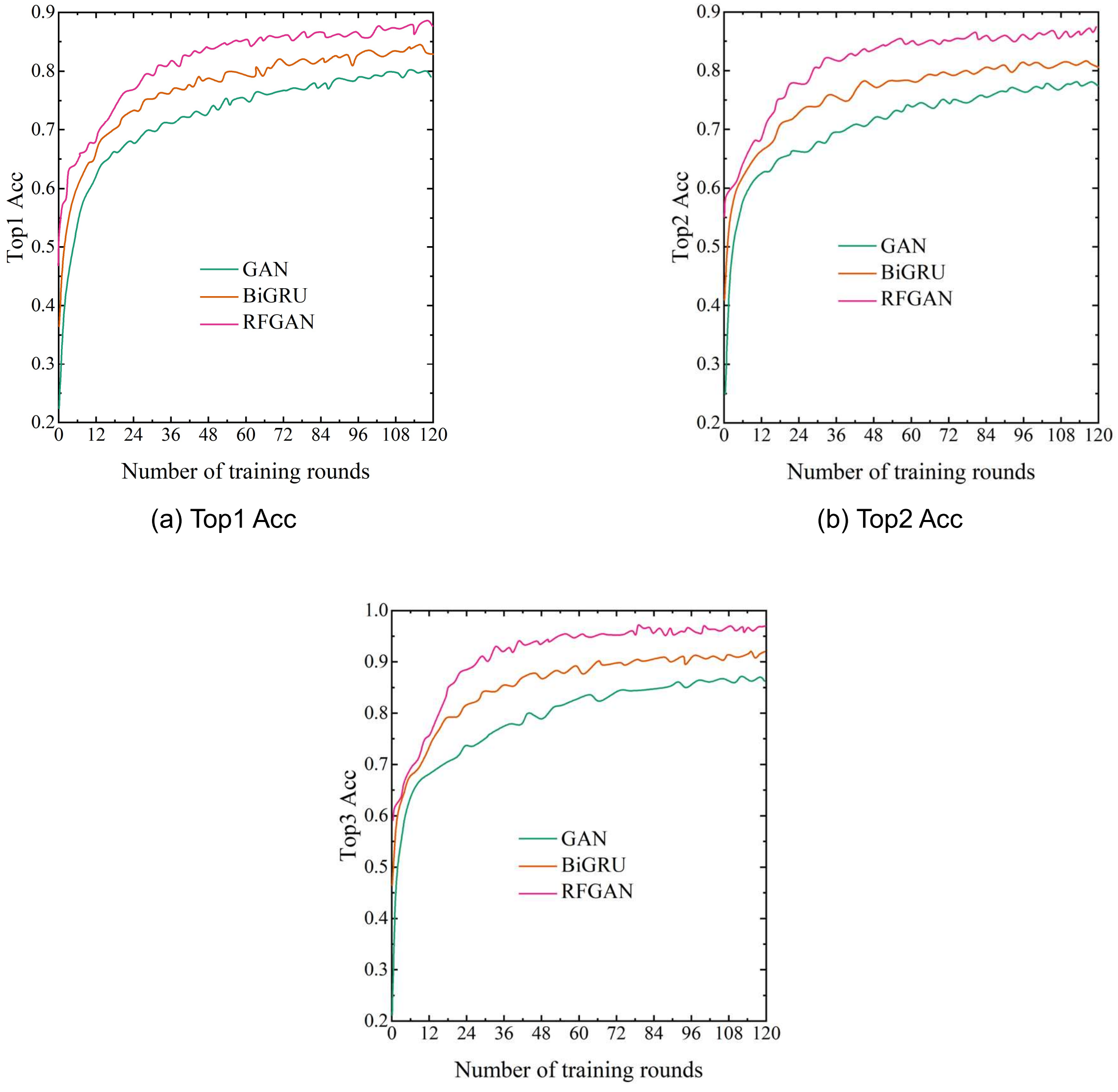
The objective assessment of the effect of music creation in traditional culture, i.e., directly from the results of note prediction of the composition model, this paper chooses the accuracy rate of note prediction as the assessment standard. The overall note prediction accuracy (ACC) is the ratio between the number of correctly predicted notes and the total number of notes, in order to enrich the objective assessment index, this paper will further refine the accuracy, Top1 Acc: when the predicted maximum probability note number and the real note number are the same, the prediction is judged to be correct, Top2 Acc: when the predicted maximum top two probability note number contains the real note number, the prediction is judged to be correct. Top3 Acc: When the predicted maximum top three probability note number contains the real note number, the prediction is judged to be correct.

GAN and BiGRU are chosen as comparison models, and the prediction results under different indicators are shown in Figure 3. Where Fig. 3(a)~(c) shows the comparison results of Top1 Acc~Top3 Acc on the validation set, respectively.

Comparing GAN and BiGRU, it can be concluded that different representations of musical features affect the final note prediction accuracy. The dimensionality of the input data chosen by both models is the same, the only difference being that the GAN inputs are note sequences represented using simple solo thermal encoding, while the BiGRU inputs are note sequences represented using the note feature vectors with contextual semantic information generated in this paper. BiGRU outperforms the results achieved by GAN, both when comparing the optimal note prediction accuracy that can be achieved by the model and when comparing the change in the note prediction accuracy curve throughout the training process.

Comparing BiGRU and RFGAN, it can be concluded that the inclusion of multiple improvements in GAN can effectively improve the final note prediction accuracy.The representation of musical features chosen by BiGRU and RFGAN is the same, which both use note feature vectors with contextual semantic information. The difference lies in the fact that the compositional model chosen by BiGRU is only BiGRU network, while the compositional model chosen by RFGAN is a combination of generator, temporal model, feature extractor, etc. based on GAN network with cyclic structure. From this, it can be known that the improved method in this paper can mine the significant information between different notes in the note sequence that is favorable for note prediction, which highlights the important information between notes and ignores the secondary information by giving different weights to the notes in the note sequence, so that the final result is a better representation of the prediction of the note sequence.

Therefore, the note prediction accuracy of the music composition model proposed in this paper reaches the highest 88.79%, 87.43%, and 96.42% on Top1 Acc, Top2 Acc, and Top3 Acc, respectively, and from the change curve of the note prediction accuracy, RFGAN also outperforms GAN and BiGRU in the whole process of training, which also verifies the proposed music composition model is effective, and the accuracy of note prediction can basically meet the requirements for use in the field of music composition.



(a) Top1 Acc

(b) Top2 Acc

(c) Top3 Acc

Figure 3: The contrast of creative effects

IV. A. 2) Comparison of Twelve Mean Rhythms

In order to be able to analyze the model-generated data against the real data, this paper also uses the twelve-mean law to count the note distributions of the generated samples and the real music data. Twelve equal temperament is a generalized method of musical law that divides a set of octave intervals into twelve equal parts equally proportional to their frequency. By counting the twelve equal-tempered note distributions of real music and generated music, it is possible to measure whether the generative model can produce a score that matches the note distribution of real music. In the generated data and real music data, 500 pieces of music are randomly selected and their twelve mean law note distributions are counted, and Fig. 4 shows the note frequency distribution.

The note distribution of the music data generated by the RFGAN model is similar to that of the training music dataset, with D, A, and B notes appearing the most frequently, while F#, G#, and C# notes all appear less frequently. It indicates that RFGAN effectively learns the note distribution law in the dataset, the note distribution basically matches, and the generated music has a specific music style.

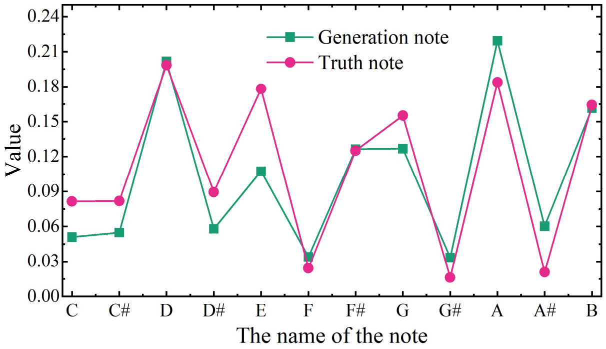


Figure 4: Note frequency distribution map

IV. B. Comparison of Subjective Evaluation of Music Composition

IV. B. 1) Analysis of musical harmony

After carrying out the objective analysis of music creation, this paper evaluates the subjective effect of RFGAN modeling for music creation by inviting people with music knowledge to verify the musical effect of the creation. While controlling the generation of music style creation, this paper evaluates whether the melodies between different tracks are harmonized or not. The experiment fine-tuned the generator so that the generated music was constantly close to a specific style of music, but this also changed the characteristics of the music, such as the harmony, so it was necessary to utilize the harmony discriminator to ensure the harmony of the music generated after fine-tuning. Since harmonized music has similar chord progressions across tracks, chord similarity is used to indicate the harmony between multiple tracks. In addition, a comprehensive analysis was performed for the rhythmic and overall effect of the music composition. Three pieces of music composed by RFGAN, GAN and BiGRU, respectively, were selected for harmonic analysis and their specific results are shown in Table 1. In the table, w/o B, w/o S, and w/o G indicate the lack of bass, strings, and guitar, respectively.

As can be seen from the table, the harmony score of the RFGAN model designed in this paper on the three pieces of music created is 2.064, which is 19.44% and 6.67% higher than that of the GAN and BiGRU models, respectively, which fully demonstrates that the music created by this paper's model possesses better harmony. In addition, the tempo and overall effect scores of the music created by this paper's model are 3.195 and 3.742, respectively, which are better than the results obtained from the comparison models. Therefore, the optimization of GAN using cyclic feature generator, reward and punishment mechanism, temporal model and feature extractor can improve the temporal feature extraction in the process of music creation, as well as improve the effect of music creation in traditional culture, and provide technical support for the promotion of the diversity of music creation.

Table 1: Analysis of Musical harmony

Model	3	w/o B	w/o S	w/o G	Rhythm	Total
GAN	1.728	1.865	1.841	2.009	2.983	3.572
BiGRU	1.935	1.942	1.976	2.115	3.064	3.615
RFGAN	2.064	2.138	2.035	2.218	3.195	3.742

IV. B. 2) Subjective evaluation of music

The evaluation of music is a complex process involving subjective feelings. People's feelings towards musical works are often influenced by a variety of factors. Therefore, in order to better evaluate the quality of model-generated music, this paper also designed a human hearing experiment to assess the model-generated music. In the experiment, we recruited 30 music enthusiasts with some music theory as volunteers, including 15 men and 15 women. When conducting the experiment, the volunteers rated the music pieces according to each of the following 4 metrics. Each model generates 15 30-bar pieces of music, and each volunteer will listen to and rate 3 pieces of music selected from each model as well as from each dataset (i.e., a total of 12 pieces of music per person).

(1) Authenticity, does this music conform to human compositional conventions?

(2) Structural, does this music have a distinct structure or repeated segments?

(3) Harmony, does the melody of this music sound smooth and harmonious?

(4) Accuracy, is this music free of compositional and performance errors?

The subjective evaluation scoring in this paper is based on a five-point rating scale. Where a score of 2 indicates excellent music quality, 1 indicates good music quality, 0 indicates average music quality, -1 indicates poor music quality, and -2 indicates very poor music quality. After the volunteers' scores for the music generated by each model were summarized and averaged, the scores for each model are shown in Table 2.

As can be seen from the data in the table, compared to other models, the RFGAN proposed in this paper performs better on all human subjective evaluation metrics, which demonstrates that RFGAN can generate harmonious and natural music. In particular, the structural scores are improved compared to other models, which indicates that the model proposed in this paper has more significant advantages in generating themes and verifies the effectiveness of the model improvement.

Table 2: Subjective Evaluation Score Form of Music

Model	Truth	Structure	Harmony	Accuracy
GAN	0.63	0.38	0.45	0.69
BiGRU	0.95	1.04	0.99	1.12
RFGAN	1.16	1.36	1.21	1.35
Truth data	1.26	1.45	1.37	1.51

V. Conclusion

The RFGAN model proposed in this paper demonstrates significant advantages in several aspects of music composition, especially in the accuracy of note prediction, with a Top1 accuracy of 88.79%. Compared with other comparative models (e.g., GAN and BiGRU), the improved model of RFGAN effectively enhances the coherence between notes and the harmony of multi-track music through the cyclic structure generator and temporal model. Through the comparative analysis of the frequency distribution of notes and the twelve equal temperament, the results show that RFGAN is able to better capture the note distribution patterns in music, and the generated music is stylistically consistent with the training dataset.

In addition, through subjective evaluation experiments, RFGAN outperforms the comparison model in terms of harmony, structure, rhythm and accuracy, which further validates its usefulness in music creation. In particular, RFGAN performs far better than traditional music generation methods in multi-track track coordination and harmony analysis, demonstrating its potential in generating compositions with complex musical features.

Through the research in this paper, it can be seen that the music creation model based on generative adversarial networks has a wide range of application prospects, especially in the fields of pop music, movie and television music and game background music, RFGAN provides a new way of thinking for automated music creation.

References

[1] Sopa, M. (2018). Local wisdom in the cultural symbol of Indonesian traditional house. KnE Social Sciences, 524-531.

[2] Yankuzo, K. I. (2014). Impact of globalization on the traditional African cultures. International Letters of Social and Humanistic Sciences, (04), 1-8.

[3] Sibani, C. M. (2018). Impact of Western culture on traditional African society: Problems and prospects. UNIZIK Journal of Religion and Human Relations, 10(1), 56-72.

[4] LUO, J., & CHEN, F. (2016). Preservation of traditional culture in modern society: A case study of China Meishan Cultural Park. International Journal of Sustainable Development and Planning, 11(3), 416-425.

[5] Carugno, G. (2018). How to protect traditional folk music? Some reflections upon traditional knowledge and copyright law. International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 31(2), 261-274.

[6] Chan, C. S. C. (2018). Sustainability of indigenous folk tales, music and cultural heritage through innovation. Journal of Cultural Heritage Management and Sustainable Development, 8(3), 342-361.

[7] Su, Z. (2025). The Influence of Religious Culture on the Style and Concepts of Traditional Chinese Music. Cultura: International Journal of Philosophy of Culture and Axiology, 22(3).

[8] Sung, C. (2021). Creating our culture together: the contemporary audience for traditional Korean music. In Routledge Handbook of Contemporary South Korea (pp. 191-208). Routledge.

[9] Leung, B. W. (2018). Traditional musics in the modern world (pp. 1-9). Springer International Publishing.

[10] D’Agostino, M. E. (2020). Reclaiming and preserving traditional music: aesthetics, ethics and technology. Organised Sound, 25(1), 106-115.

[11] Hogenes, M., Oers, B. V., & Diekstra, R. F. (2014). Music composition in the music curriculum. US-China Education Review A, 4(3), 149-162.

[12] Onyeji, C. U. (2016). Composing art music based on African indigenous musical paradigms. An Inaugral lecture of the University of Nigeria Nsukka,(102nd), publ. by University of Nigeria Senate Ceremonials Committee.

[13] Lizeray, J. Y. M., & Lum, C. H. (2019). Semionauts of Tradition: Music, Culture and Identity in Contemporary Singapore. Springer.

[14] Moir, Z., & Medbøe, H. (2015). Reframing popular music composition as performance-centred practice. Journal of Music, Technology & Education, 8(2), 147-161.

[15] Ning, H., & Maneewattana, C. (2024). The characteristics and forms of contemporary Chinese Zheng music composition. Journal of Roi Kaensarn Academi, 9(3), 405-419.

[16] Willgoss, R. (2012). Creativity in contemporary art music composition. International Review of the Aesthetics and Sociology of Music, 423-437

[17] Lerdahl, F. (2019). Composition and cognition: Reflections on contemporary music and the musical mind. Univ of California Press.

[18] Li, N., & Chen, Y. (2024). Influence of traditional Chinese thought on the performance style and creation of piano music works. Trans/Form/Ação, 47(5), e02400155.

[19] Xu, N. (2018). Analysis of the Correlation Between Folk Music Education and Chinese Traditional Culture. Educational Sciences: Theory & Practice, 18(5).

[20] Zhang, Y., Zhou, Z., & Sun, M. (2022). Influence of musical elements on the perception of 'Chinese style'in music. Cognitive Computation and Systems, 4(2), 147-164.

[21] Andrijauskas, A. (2016). Visual Arts and Music in Traditional Chinese Art System. Music in Art, 41(1-2), 165-187.

[22] Li, J., & Heng, T. (2023). The creative thinking deduction of traditional Chinese piano music elements-taking cultural works from different periods in history as the main line. Herança, 6(1), 244-256.

[23] Linwei Li. (2024). Teaching Research on the Integration of Piano Playing Technique and Music Theory Knowledge. Research and Commentary on Humanities and Arts,2(12).

[24] Saebyul Park,Eunjin Choi,Jeounghoon Kim & Juhan Nam. (2024). Mel2Word: A Text-Based Melody Representation for Symbolic Music Analysis. Music & Science,7.

[25] Tarannum Shaikh & Ashish Jadhav. (2024). Music Generation Using Dual Interactive Wasserstein Fourier Acquisitive Generative Adversarial Network. International Journal of Computational Intelligence and Applications,24(01).