International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-12

https://doi.org/10.70917/ijcisim-2026-0040
Article

A Study of Music Composition Models Combining Generative Adversarial Networks and Their Application to Film Scoring

Jiangli Jia * and Juan Chen [2]

[1] Department of Music, Shanxi University, Taiyuan, Shanxi, 030006, China
[2] School of Art, North University of China, Taiyuan, Shanxi, 030051, China;
* Correspondence author: yyr200203@163.com

Abstract: Automatic generation of movie music is one of the current research hotspots in the field of artificial intelligence. This study proposes a music composition model based on the application of artificial intelligence in movie soundtrack composition. The model is based on generative adversarial network and its derivative model CT-GAN, and builds the MTC-GAN music generation method through multi-track correlation modeling, temporal structure modeling and discretization processing. In the objective index comparison experiments, the experimental indexes verify that the MTC-GAN model in this paper can create high-quality music, and its Scale Consistency, Tone Span, and Uniqueness structures are 89.13%, 13.84, and 64.43, respectively, which effectively improves the quality of music generated by the music composition model. The model-generated movie soundtrack score has an overall improvement of 1.48% over the human work score, and its evaluation scores in terms of melody, rhythm and emotion are optimal among all compared models. Experiments show that the method in this paper achieves ideal results in generating music quality, which is helpful to help movie soundtrack creation and promote the innovation and development of movie soundtrack art.

Keywords: generative adversarial network; multi-track; music creation; CT-GAN; movie soundtracks

1. Introduction

In the new period, with the advancement of all kinds of new technologies, the development of various industries has been greatly affected, and the overall business efficiency and business quality have been improved, which is also true in the development of the movie industry. In the actual development of the film industry, not only do we need new technologies to ensure the efficiency of film production and publicity, but we also need to make full use of new technologies to complete the corresponding tasks, including shooting, editing, soundtracks and other auxiliary work [1-3].

Movie soundtrack is an indispensable part of movie art, which can not only enhance the emotional expression of the movie, but also create a unique atmosphere, so that the audience is more deeply immersed in the storyline of the movie [4-6]. And for those movies that are quite unique, embody the characteristics of the times, and are full of imagination and sci-fi, computer music technology has played a great role in promoting the development of their music [7]. Traditional movie soundtracks need real orchestras to cooperate with the completion, which will increase the investment of the film, and a lot of film shooting can only be completed by the stronger film companies, thus greatly limiting the development of the film industry [8-9]. Computer music technology, on the other hand, only requires a computer, a music arranger and related audio equipment, which can easily complete the soundtrack of a movie [10-11]. At the same time, computer music technology provides a broader creative space for filmmakers, who can freely carry out video production according to their own imagination, and as a

result, many excellent movie works have appeared [12-14]. It can be said that a high-level movie soundtrack is absolutely inseparable from the development of audio technology, which is an important prerequisite for the development of movie music.

The article discusses the application of artificial intelligence in film soundtrack composition, based on the CT-GAN network model, analyzes the inter-track correlation, proposes a temporal structure model to satisfy the temporal characteristics of the music, and uses the hard thresholding method to discretize the output results, and proposes a music composition model based on MTC-GAN. MIDI music samples collected from the network are used as experimental objects to carry out training and testing of the model. The model complexity of SR-CNN-VAEGAN, Music VAE and the method in this paper are compared. Subsequently, the evaluation discovery of different model-generated music samples is carried out in terms of several objective evaluation metrics, such as Scale Consistency, Tone Span and Uniqueness, to explore the performance of the multi-track music composition model in this paper. Finally, the comparison between model intelligent film score composition and artificial film score composition is carried out, and two groups of professional composers and non-composers are selected for scoring, and the scoring results of different models are analyzed based on the four aspects of Melody, Harmony, Rhythm, and Emotion to explore the effect of the MTC-GAN model in film score composition.

2. Application of Artificial Intelligence in the Creation of Movie Soundtracks

With the rapid progress of science and technology, artificial intelligence is gradually penetrating into various fields of film art, of which the field of film soundtrack creation and production is particularly significant. The application of artificial intelligence not only brings unprecedented convenience and possibilities to this traditional art form, but also profoundly promotes the overall development of movie art.

2.1. Quick Generation of Music Samples

With the rise of AI tools, the field of movie music creation has ushered in a new mode of collaboration. Under this model, composers are able to guide AI to quickly generate music samples and directly participate in the creative process, achieving instant synchronization and feedback between music and visual content. In the creative germination stage, the AI tool can intelligently generate matching music, scene previews, sound effects and vocals according to the textual description of the script, building a preliminary audiovisual conceptual framework for the creative team. Entering the virtual preview stage, AI tools, such as generating adversarial network models, can help the music department to quickly produce music samples and make real-time adjustments based on the director's feedback. This efficient way of working enables all staff members to deeply understand the director's creative intent and direction from the perspective of the audiovisual whole.

2.2. Assist in Identifying Creative Solutions

With the AI tool, the composer can input to the AI system a detailed description of the movie, the emotional requirements of the scene, as well as his personal compositional style and preferences. Based on these inputs, the AI system utilizes deep learning models (e.g., LSTM networks) to quickly generate a variety of creative musical scenarios. These scenarios may contain different melodic directions, chord progressions, instrumental configurations, and sound elements to reflect different emotions and atmospheres in the movie. The composer can select one or more of these solutions that he thinks best matches the emotion of the film as a basis, and make further modifications and optimizations on that basis to create a musical composition that perfectly fits the content of the film, a process that greatly broadens the boundaries of creation and shortens the conceptualization time. In addition, AI can analyze market trends and audience preferences, predict the market acceptance of music design, and provide composers with data-supported decision-making basis. This intelligent assistance not only speeds up the music design process, but also ensures that the music fits the content of the movie, improving overall production efficiency and quality.

2.3. Increased Efficiency of the Music Program

With its powerful data processing capability and innovative creation assistance functions, artificial intelligence tools can not only efficiently complete the basic work in arranging, such as the optimization of instrumental combinations and the conception of harmonic progression, but also replace manpower to a certain extent in arranging complex and time-consuming musical elements, thus significantly shortening the production cycle and enhancing work efficiency. By analyzing a large number of musical

works and mastering the laws of harmony between different instruments, Generative Adversarial Network Model is able to automatically generate an arrangement that meets the requirements, which is undoubtedly a cost-effective choice, especially for film projects with limited budgets.

3. MCT-GAN Based Music Composition Modeling

Based on the application of artificial intelligence networks in movie soundtrack composition, this paper utilizes generative adversarial networks to construct a music composition model. The model uses CT-GAN, a GAN-derived model with better performance, as the generative adversarial model in music generation.

3.1. CT-GAN Model

(1) Generative Adversarial Networks

Generative Adversarial Network (GAN) is an unsupervised learning method that expands minority class samples by training generators to generate generated data that approximate the real data distribution.

Two models (game parties) are included in the framework of Generative Adversarial Networks: the discriminator D, which is used to identify whether a sample belongs to the real data, and the generator G, which is used to learn the distribution of the real data and generate new data samples. Both sides of the game achieve their respective victories by continuously optimizing their capabilities, which will ultimately produce a Nash equilibrium between the two sides. The input to the generator is random noise z, the noise passes through the generator G to produce new samples $G(z)$, and the generated samples obey the distribution p_G, while the input to the discriminator is the real data samples x as well as the generated samples $G(z)$, and the real samples obey the distribution p_{data}. The goal of the discriminator is to identify the true data sample x as true and the generated sample $G(z)$ as false as correctly as possible. The generator's goal is to produce generated samples that are as close as possible to the true sample distribution p_{data}, so that it can fool the discriminator D. Both sides through the continuous confrontation to optimize their own capabilities, so that the performance of the model is gradually improved, and ultimately, because the generator to generate samples of the distribution p_G has been very close to the distribution of the real data p_{data}, the discriminator D will not be able to identify the sample class, at this time, the model training is complete, the generator G has been able to generate very real data samples.

(2) CT-GAN

CT-GAN is an algorithm based on the method of GAN and optimized for tabular data, through the continuous training of the generator and discriminator, it can effectively generate discrete and continuous data in the table, CT-GAN is designed to design a new way of normalizing continuous feature columns and a conditional trigger for the training of imbalanced discrete feature columns sampling relative to GAN.

First, in order to generate both discrete and continuous features in tabular data, the algorithm needs to use softmax function and tanh function in the output layer. Among them, the softmax function maps the output result to the interval from 0 to positive infinity through the exponential function, so that the output value is non-negative, and then transforms the processed model output category results into the form of probability through normalization. It is calculated as in equation (1):

$$softmax(f)_y = \frac{e^{(f_y)}}{\sum_{c=1}^{C} e^{(f_c)}} \tag{1}$$

where f_y is the prediction result of the y th category of the classification problem, f_c is the prediction result of the c th category, and C is the total number of categories. softmax converts the output into the probability of belonging to each category, which is between 0 and 1. The tanh function maps the output to the -1 to 1 range through exponential function processing and min- max normalization. max normalized mapping to between -1 and 1, which is calculated as in equation (2):

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2}$$

where x is the output result of the classification problem. However, when the output layer uses the tanh function to output the result, there may be a multi-peaked distribution, and the min-max normalization used in tanh will easily lead to the problem of vanishing gradient arises. As a result, a distribution-specific normalization is designed in CT-GAN to overcome the non-Gaussian distribution, which will process each column of feature data in the data individually, and each value C_i in each feature $C_{i,j}$ will be in the form of a unique thermal encoding of the specific distribution it belongs to in the multipeak distribution. The steps are: for each continuous type of feature column, a Gaussian Mixture Model (GMM) is used to estimate the number of single distributions in it and fit a Gaussian Mixture Model. η_i denotes the i th single distribution in the multimodal distribution, and then the probability of each sample coming from each specific single distribution in the multimodal distribution is calculated, and assuming that the probability of a sample coming from the third single distribution is the largest, the distribution is denoted by a uniquely hot coding of the form $\beta_{i,j} = [0,0,1]$, and using a scalar $\alpha_{i,j}$ to denote the value of the sample in that single distribution, and $\alpha_{i,j}$ is computed as in Equation (3):

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_3}{4\phi_3} \tag{3}$$

where η_3 and ϕ_3 denote the weight and standard deviation of the third single distribution, respectively. Ultimately, a single row of data samples will be represented by the scalar $\alpha_{i,j}$ together with the uniquely hot coding $\beta_{i,j}$ and the uniquely hot coding of the discrete feature columns.

Second, in order to uniformly sample each category in the discrete feature columns and to restore the distribution of the real samples in the subsequent tests, a conditional trigger for unbalanced discrete feature column sampling training is designed in CT-GAN. The implementation of the conditional trigger will be based on three components: the conditional vector, the generator loss, and the sampling training method. Among them, the condition vector $cond$ is used to represent the current conditions, i.e., one of the categories of the discrete feature columns is represented using the unique hot coding, such as there are two categories in the discrete feature columns and the sample belongs to the first, it is denoted as [1,0], and multiple discrete feature columns are merged by direct splicing, e.g., the sample's feature one contains three categories, and two categories are contained in the feature two, and the generator is specified as feature two is [1,0] of the sample, the condition vector $cond$ can be expressed as [0,0,0,0,1,0].

Finally, the sampling training method, on the other hand, samples the conditional vector uniformly with the training data to achieve an effective exploration of the possible class values in the discrete feature columns, and then evaluates the outputs produced by the conditional triggers by estimating the distance between the conditional distributions of the data generated by the generator and the conditional distributions of the real data to assess the goodness of the results.

3.2. MCT-GAN Model

Based on the CT-GAN model, the complete MTC-GAN model is constructed through multi-track correlation modeling, temporal structure modeling, and discretization of the generated results, and then integrated.

3.2.1. Multi-Track Correlation Modeling

The music creation methods are categorized as “jamming” and “composer” depending on the way the tracks are presented. By combining the ideas of Jamming and Composer, a hybrid (Hybird) model can be further proposed that utilizes both intra-track and inter-track discriminative feedback.

The Hybird model uses M generators to create music for M tracks, each taking as input the inter-track random vector z and the intra-track random vector z_i. At the same time, it is expected that

the inter-track random vectors can coordinate the generation of different music (G_i), so that only one discriminator is used to jointly evaluate M tracks. The Hybird model is used in the MCT-GAN of this paper.

3.2.2. Time-Structured Models

"Generate from scratch" (T1) means generating music clips from scratch for multiple tracks simultaneously. The generator G is divided into two main parts, G_{tcmp} represents the time structure generator and G_{bar} represents the bar generator:

$$G^{(t)}(z_t) = G_{bar}(G^{(t)}{}_{temp}(z_t)) \ \forall t = 1,2,...,T \tag{4}$$

where z_t is a time-dependent random vector and T is the number of bars contained in each phrase. G_{temp} takes z_t as input, and for each bar generates a latent vector carrying temporal information, which is then used by G_{bar} to sequentially generate music. Over time, the set of $G^{(t)}(z_t)$ becomes a tensor.

Conditional generation (T2) for a given track means that the music of the other tracks is created based on the music of one of the tracks given in advance, thus generating the whole song. The generator G sequentially generates the music bars of the other tracks using the conditional generator $\vec{G}_{bar}$ in order. The $\vec{G}_{bar}$ has two inputs, random noise z and a bar $\vec{x}^{(t)}$ of a given track, and generates a bar based on that track. Moreover, $\vec{x}^{(t)}$ is projected to the space of z by encoder E, which is then connected to z as the input to the generator, where encoder E and generator G and discriminator D are all architectures composing different convolutional neural networks (CNNs), respectively:

$$G^{(t)}(z_t) = \overrightarrow{G}_{bar}(z^{(t)} \circ E(\vec{x}^{(t)})) \forall t = 1,2,...,T \tag{5}$$

where $\circ$ denotes the connection of vectors.

In this paper, we propose and use a temporal structure modeling method that combines the two (T3).The basic idea of T3 is: firstly, according to the idea of T1, the music of a specific track is generated, then the music of the generated track is inputted to the encoder E, and through E, the track is projected into the space of z, and then after connecting them as the input of the generator, the generator is allowed to learn the temporal structure under the track and complete the song by generating the music for the remaining tracks in order accordingly.

z_t are time-dependent random vectors, z_t is the input of the temporal structure generator G_{temp}, and the potential vector G_{temp} output carrying time information $G_{temp}(z_t)$ or $\hat{z}$ is used as the input of the bar generator G_{bar} to generate music for a track of music $G_{bar}(\hat{z})$ or $\vec{x}$. Next, the encoder E projects $\vec{x}$ into the space of random noise z and connects them as input to the generator G:

$$\vec{x}^{(t)} = G_{bar}(G_{temp}(z_t)^{(t)}) \forall t = 1,2,...,T \tag{6}$$

$$G^{(t)}(z) = \overrightarrow{G}_{bar}(z^{(t)} \circ E(\vec{x}^{(t)})) \quad \forall t = 1,2,...,T \tag{7}$$

3.2.3. Discrete Processing

Deep Convolutional Generative Adversarial Networks can only generate real-valued pianoroll format, so to achieve the generation of music, the generation results need to be further discretized to obtain the final binary-valued results. In MCT-GAN, a hard thresholding method is used to discretize the generation results, the basic idea is: the last transposed convolutional layer of the generator uses the Tanh function as the activation function, takes the output value in the range of (-1,1), and then divides the output value into: values greater than 0 and values less than 0 by setting the threshold value to zero. For the part greater than 0, black color is taken to indicate that a note is played at the current position, and for the part less than 0, white color is taken to indicate that no note is played at the current position.

3.2.4. MCT-GAN

The network model of the complete MCT-GAN is obtained by combining and extending the multi-track correlation model, the time structure model and the discretization processing method mentioned above.

The network is mainly divided into two parts: the first part generates music for a specific track, G_{temp} with $z_t^{'}$ as input, G_{bar} with G_{temp} as the output of $G_{temp}(z_t^{'})$ (or $\hat{z}$) as input, and then outputs the music of one track $G_{bar}(\hat{z})$, or $\hat{x}$:

$$\hat{x}^{(t)} = G_{bar}(G_{temp}(z_t^{'})^{(t)}) \forall t = 1,2,...,T \tag{8}$$

The second part contains four parts of inputs: inter-track time-independent random vector z, intra-track time-independent random vector z_i, inter-track time-dependent random vector z_t, and intra-track time-dependent random vector $z_{t,i}$. First, the inter-track time-independent random vector z and the intra-track time-independent random vector z_i are concatenated to obtain $\hat{z}_i$, respectively, and the inter-track time-dependent random vector z_t and the intra-track time-dependent random vector $z_{t,i}$ are concatenated to obtain $\hat{z}_{t,i}$. After concatenation, it is fed into generator G together with $E(\hat{x}^{(t)})$ generated in the first part and mapped by decoder E to generate the music for the other tracks in turn:

$$G_i^{(t)}(\overline{z}) = G(\hat{z}_i \circ \hat{z}_{i,t} \circ E(\hat{x}^{(t)})) \ \forall t = 1,2,...,T \tag{9}$$

Finally, the discriminator D is trained using the G-generated data along with the real data $\tilde{x}$. Since MCT-GAN uses CT-GAN as its generative adversarial network model, then a consistency penalty term (CT) needs to be added to the objective function of MCT-GAN.

Denote by $d(a,b)$ the l_2 distance between a and b, which need not be penalized if for the discriminator $D: x \mapsto y$ if there exists a constant $M \geq 0$ such that any $x, x^{'} \in X$, satisfies inequality (10). If the inequality cannot be satisfied, such cases need to be penalized:

$$d(D(x), D(x^{'})) \leq M \cdot d(x, x^{'}) \tag{10}$$

The consistency penalty rule described above can be realized by adding the following consistency penalty term to the objective function:

$$CT|_{x,x^{'}} = E_{x,x^{'}}\left\{\theta\left[\max\left(0, \frac{d(D(x), D(x^{'}))}{d(x, x^{'})} - M^{'}\right)\right]\right\} \tag{11}$$

In MCT-GAN, the hidden layer of the discriminator is perturbed using the dropout method instead of perturbing the input x. When the dropout ratio is small, the output of the perturbed discriminator can be considered as the output of the purified discriminator as a response to a "virtual" data point $x^{'}$ not far away from x. Thus, after applying dropout to the hidden layer, the output of the discriminator is represented by $D(x^{'})$. In the same manner, the (random) dropout is again applied to the hidden layer of the discriminator, so that the second virtual point $x^{'}$ around x is found and the corresponding output is represented by $D(x^{'})$. However, computing the distance $d(x^{'}, x^{'})$ between two virtual data points is not possible, where it is assumed to be bounded by a constant and this constant is denoted as $M^{'}$.

Thus, the final consistency regularization formula is as follows:

$$CT|_{x^{'}x^{'}} = E_{x \sim p_r}[\max(0, d(D(x^{'}), D(x^{''})) + 0.1 \cdot d(D\_(x^{'}), D\_(x^{''})) - M^{'})] \tag{12}$$

The weights of the discriminators were updated using Eq. (13) with the added CT penalty term as the MCT-GAN objective function:

$$L = E_{z\sim P_z}[D(G(z))] - E_{x\sim P_r}[D(x)] + \lambda_1 GP|_{\hat{x}} + \lambda_2 CT|_{x',x'} \tag{13}$$

4. Experimental Analysis of Music Composition Models

4.1. Data Sets

The dataset used for the experiments was a total of 1000 single-track monophonic MIDI music samples collected on the web. The MIDI data was normalized and cut into fixed-length music samples, which were then encoded and converted into two-dimensional sequences for input into the model for training.

4.2. Analysis of Experimental Results

4.2.1. Model Training

The negative penalty discrimination loss of the discriminator in the MTC-GAN model is negatively correlated with the generation quality of the model as well as the degree of convergence, i.e., the smaller the value of the loss, the better the quality of the generated samples. In this paper, the TC-GAN model is trained to compare with the MTC-GAN model. The negative penalty discriminative loss values for each step in the training process of the two models are calculated, and Fig. 1 shows the comparison curves of the loss values of the models in the training of 100k steps. The MTC-GAN music composition model can get lower negative penalty discriminative loss (about 1.55), and the convergence degree of the model training is accelerated, which indicates that the MTC-GAN model can improve the quality of the model's music generation. The TC -GAN model converges slower after about 18k steps of training, starts to be able to generate regular music, and converges completely after 40k steps of training, with optimal generation.

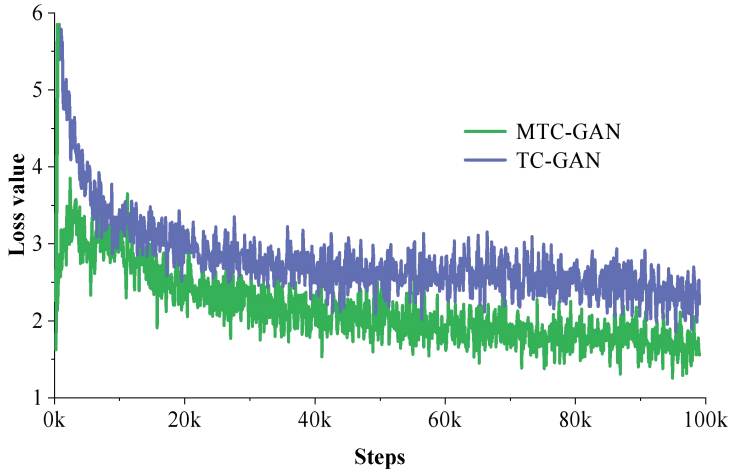

Figure 1. Loss value comparison curve of model training 100k steps.

The training speed of the model is also an important index to evaluate the performance of the model, in this paper, the current mainstream music generation models Music VAE and SR-CNN-VAEGAN are selected to compare the complexity with MTC-GAN respectively. In the experiments, each model is trained five times using the same dataset and the results are averaged. Table 1 shows the complexity comparison of each model, and the number of parameters of MTC-GAN model is lower than that of other models, which is 2.18 Million, and there is a certain degree of improvement both in the number of convergence rounds (280 epochs) and the training time (6,600 s) compared with other models.

Table 1. Comparison of the complexity of the model.

Models	SR-CNN-VAEGAN	Music VAE	MTC-GAN
Parameter quantity (Million)	11.89	4.52	2.18
Convergence number(epoch)	320	610	280
Training time(s)	63975	25025	6600
Unit time (s)	218	44	25

4.2.2. Comparison of Objective Evaluation Indicators

In order to be able to evaluate the music generation effect of the MTC-GAN model more objectively, Scale Consistency (SC) was defined to count the percentage of notes in the generated samples that could best match the standard scale, Tone Span (TS) was used to calculate the number of semitone steps between the lowest and the highest notes in the samples, and Uniqueness (UN) was used to calculate the percentage of all notes played uniquely once in all time steps.

Figure 2 shows the evaluation metrics curve of 200 randomly generated movie soundtrack music samples after the training of the MTC-GAN model is completed, and the fluctuation of the metrics of the generated samples proves the diversity of the movie soundtracks generated by MTC-GAN.

In this paper, we also choose several types of mainstream music generation models for comparison tests, the comparison models are SR-CNN-VAEGAN, Music-VAE and TC-GAN models, using the same dataset to train the models and generate music. From the real music data and the generated music data of each model, 200 pieces of 5-bar (10-second) music are randomly selected, and the statistical averages of the samples from the same source under different evaluation indexes are computed for comparison, and the results of the comparison of the objective evaluation indexes are shown in Table 2.The average value of the Scale Consistency of the music generated by the MTC-GAN is about 89.13%, which is close to the real one, and the average value of the Scale Consistency is 90%. The average value of Scale Consistency for music is 90.56%, which is higher than 87.26% for SR-CNN-VAEGAN, 80.62% for MusicVAE and 87.48% for TC-GAN. As for Tone Span and Uniqueness, the average values of the movie soundtrack samples generated by MTC-GAN are 13.84 and 64.43, respectively, which are very close to the values of 14.21 and 63.88 for the two metrics of real music, which are better than the other comparative models, suggesting that the music generated by MTC-GAN is more similar to real music in terms of note uniqueness and pitch span.

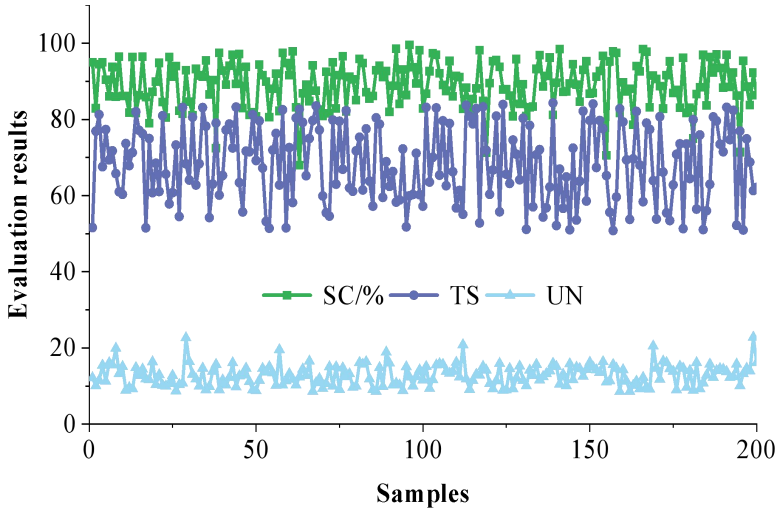

Figure 2. The evaluation results of the MTC-GAN model generation sample.

Table 2. Comparison of objective evaluation indicators.

Models	Scale Consistency/%	Uniqueness	Tone Span
SR-CNN-VAEGAN	87.26%	38.22	18.46
Music VAE	80.62%	65.43	31.79
TC-GAN	87.48%	61.55	12.96
MTC-GAN	**89.13%**	**64.43**	**13.84**
REAL MUSIC	90.56%	63.88	14.21

4.2.3. Comparison of Twelve Mean Rhythms

In order to be able to analyze the comparison between the model generated data and the real data, this paper also used the twelve-mean law to count the note distribution of the generated samples and the real music data. In the generated data and the real music data, 200 pieces of music were randomly selected and their twelve mean law note distribution was counted, and Figure 3 shows the results of the note frequency distribution. The music data generated by the MTC-GAN model is similar to the note

distribution of the training music dataset, with C#, F#, G#, and A notes appearing the most often, all with distributions above 0.1, and D#, A#, and B notes all appearing at lower frequencies, all below 0.05, which indicates that the MTC-GAN model effectively learns the note distribution law in the dataset, and the note distributions basically coincide with each other, and the generated music with a specific movie soundtrack music style.

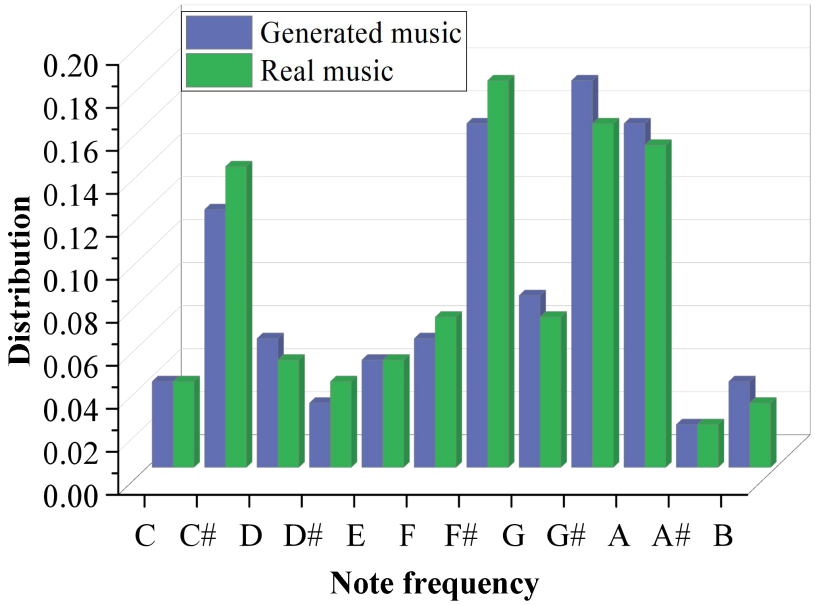

Figure 3. Distribution of notes frequency.

4.3. Generating Movie Soundtrack Evaluations

Split all participants into two groups, professional composers and non-composers. Participants in the professional group were those with educational degrees in music composition or electronic music composition and production.

4.3.1. Human Composition & Intelligent Composition

First, a mixed music collection of five movie soundtrack pieces by professional human composers and five movie soundtrack pieces by the models in this paper was prepared for people to determine whether they were composed by humans or by the AI.30 professional composers were asked to rate each piece of music they heard from the theoretical side of music composition, while 50 non-composers were asked to rate each piece of music based on their subjective perception of it scoring. Each listener would evaluate and score the test sample (on a scale ranging from 1 to 10), summarizing the final scores. In order to avoid participants being influenced by other aspects, such as instrumental timbre, all test music, whether composed by humans or models, was derived from the same instrument.

The results of the evaluation of human compositions and AI compositions are shown in Figure 4. Among the professional composers, the average scores of the human musical compositions were higher than those of the AI compositions, which were 8.08 and 7.94, respectively. However, among all participants, the score of music composed by the MTC-GAN model was 8.22 higher than the score of real human compositions (8.10), which indicates that the quality of AI movie soundtrack compositions by the MTC-GAN model is very close to the quality of human composers.

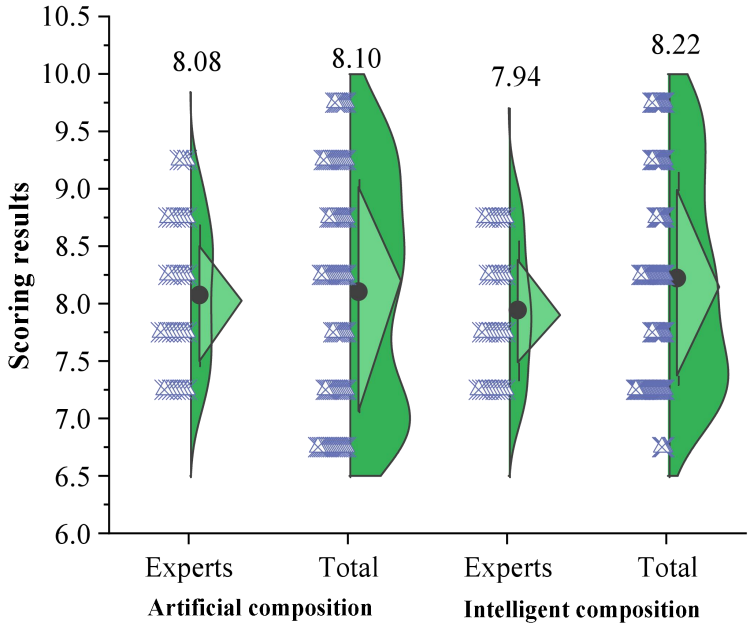

Figure 4. The evaluation results of artificial composition and artificial intelligence composition.

4.3.2. Comparison Experiments

The experiment compares the generated samples to SR-CNN-VAEGAN, Music-VAE, and TC-GAN. Participants (20 composers and 25 non-composers) were presented with 25 movie soundtrack clips from four different models, all trained on the same training set given the same start notes and instrumental timbres, and each model generated five music clips. Participants were then asked to evaluate and score the music in terms of melody, harmony, rhythm, and emotion, and after summarizing the scores, the mean and standard deviation were obtained. The results of the four model scores are shown in Figure 5. Compared with SR-CNN-VAEGAN, Music-VAE, and TC-GAN generation models, the overall quality of MTC-GAN model is significantly improved, except for harmony, other index scores are significantly higher than the other three models, and the overall evaluation score is 8.07, which is 12.71%~42.08% higher than that of the other models, indicating that the MTC-GAN model generated movie soundtracks are more in line with the requirements and rules of composing, and need to strengthen the research in this aspect of harmony in future work.

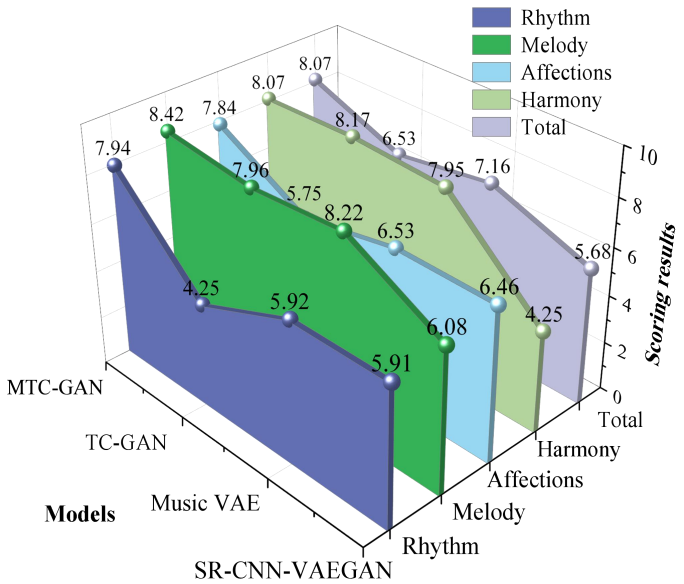

Figure 5. Scoring results of four models.

5. Conclusion

As an important part of the movie, the soundtrack is the key to express the emotion and narrative of the movie. In this paper, we propose a music composition model combined with generative adversarial network, construct a music generation method based on MTC-GAN, and conduct experimental evaluation and movie soundtrack application. The experimental results show that the MTC-GAN model in this paper has better convergence effect, and its test results on Scale Consistency, Tone Span and Uniqueness are 89.13%, 13.84 and 64.43 respectively, which are better than the comparative music generation methods and very close to the index value of real music. At the same time, the twelve-mean-tempered note distribution of the music generated by the MTC-GAN model is basically consistent with that of real music, which indicates that the proposed generative adversarial network model has a good performance in music creation. Using the model to generate movie soundtracks and comparing it with the human work score, the overall evaluation scores of the model composition and the human work score are 8.22 and 8.10, and the scores of the model-generated movie soundtracks are higher than those of the comparison method in the three aspects of melody, tempo, and emotion, which indicates its applicability and superiority in the application of movie soundtracks.

The music composition method proposed in this paper can provide more possibilities for those engaged in music composition and movie soundtrack production, bringing them more creative and inspiring inspirations, thus promoting the development of the music industry. However, there are some shortcomings, the harmony of the model-generated music is still poor, and there is room for improvement. Subsequent research will further explore ways to enhance the harmony of music, such as trying to combine the chords and other music theory rules to constrain the generation process, and further extend the model to other conditions of controllable generation, such as music emotion, playing instruments, etc. in order to enhance the human-computer interaction ability of the music generation model, and to satisfy the diverse needs of movie creation.

References

1. Alforova, Z., Marchenko, S., Kot, H., Medvedieva, A., & Moussienko, O. (2021). Impact of digital technologies on the development of modern film production and television. Rupkatha Journal on Interdisciplinary Studies in Humanities, 13(4), 1-11.
2. Salvador, E., Simon, J. P., & Benghozi, P. J. (2019). Facing disruption: The cinema value chain in the digital age. International Journal of Arts Management, 25-40.
3. Sun, P. (2024, July). Digital Optimization of Film and Television in the Era of Artificial Intelligence. In 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC) (pp. 558-562). IEEE.
4. Phetorant, D. (2020). Peran musik dalam film score. Journal of Music Science, Technology, and Industry, 3(1), 91-102.
5. Gillick, J., & Bamman, D. (2018, June). Telling stories with soundtracks: an empirical analysis of music in film. In Proceedings of the First Workshop on Storytelling (pp. 33-42).
6. Moreira, A., & Chambel, T. (2018, June). As Music Goes By in versions and movies along time. In Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video (pp. 239-244).
7. Soshalskyi, O. (2023). Modern technologies for creating film music content. Notes on Art Criticism, 2(23), 155-160.
8. Ma, B., Greer, T., Knox, D., & Narayanan, S. (2021). A computational lens into how music characterizes genre in film. PloS one, 16(4), e0249957.
9. Tan, S. L., Spackman, M. P., & Wakefield, E. M. (2017). The effects of diegetic and nondiegetic music on viewers' interpretations of a film scene. Music Perception: An Interdisciplinary Journal, 34(5), 605-623.
10. Zhang, C. (2022). Research on IMDB film score prediction based on improved whale algorithm. Procedia Computer Science, 208, 361-366.
11. Liu, Y., Zhang, M., & Zhang, M. (2021). Movie Score Predication Model Based on Multiple Nonlinear Regression. Tehnički vjesnik, 28(3), 914-921.
12. Lin, T. F., & Chen, L. B. (2024). Harmony and algorithm: Exploring the advancements and impacts of AI-generated music. IEEE potentials.
13. Xiang, Z., & Guo, Y. (2020). Controlling melody structures in automatic game soundtrack compositions with adversarial learning guided gaussian mixture models. IEEE Transactions on Games, 13(2), 193-204.
14. Wan, C. H., Chuang, S. P., & Lee, H. Y. (2019, May). Towards audio to scene image synthesis using generative adversarial network. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 496-500). IEEE.