Style Transfer of Traditional Chinese Bamboo Flute Music Based on AI Model

1st Ziqi Qin
Beijing National Day School
Beijing, China
qinziqi18@126.com

2nd Yixuan Huang
Beijing National Day School
Beijing, China
kathelin_huang@163.com

3rd Zijie Zheng
Beijing National Day School
Beijing, China
zhengzijie@bnds.cn

***Abstract*—The combination of computer technology and music can promote the development of multi-source and diversity of the music style, which can effectively promote the exchange of music culture and the enrichment of the music market. In this paper, we present a deep cycle generative adversarial networks (DNN-CycleGAN) model to complete the transformation of the music style between the bamboo flute and flute. We use a short-term Fourier transform algorithm (STFT) to realize the time-frequency conversion of the waveform music, and then utilizing a deep convolutional neural network structure (DNN) to extract the features of the waveform music in the Chinese bamboo flute. Then we structure the CycleGAN model to transfer the bamboo flute music into flute music. We collect a sufficient amount of music data sets to train our model. The experimental results show that the converted music has a precise rhythm, and the timbre characteristics of the flute and bamboo flute are demonstrated.**

***Index Terms*—music style, STFT, CycleGAN, bamboo flute, flute**

I. Introduction

Music can not only express people's emotions but also promote communication between people. The variety of musical works is helpful to promote the spread of musical culture and the development of the musical market [1], [2]. With the rapid development of modern computer technology and Internet techniques, digital synthetic music has become the trend of traditional music development. The emergence of various forms of digital music works has become the main industry in the music market by virtue of its individuality and diversity [3], [4]. Meanwhile, with deep learning technology (DLT), the musical style transition is possible.

The music style is the various musical elements in the category of music, such as tune, rhythm, timbre, intensity, harmony, texture, and form, but it mainly refers to the tune. Musical style transition is to retain the content characteristics of one piece of music from one instrument and to convert it into a piece of new music with the style characteristics of another piece of music from another instrument [5]. The waveform music produced by the performer using the instrument is nonlinear in tone, timbre, and volume, which increases the difficulty of modeling. Especially, because the waveform music data usually has a high time sampling rate, and modeling directly on the raw sampled data sets is challenging.

At present, in the waveform of music produced from the instrument, there are two kinds of methods employed to transform the musical styles: one is based on the sound signal processing ways and the other is image processing ways. The sound signal processing ways use mainly the original music sampling waveform to train the model and to realize the transformation of the musical styles in different instruments. The main models include the simple neural network model and series-based neural network model. In [6], the author proposes a WaveNet model to train the audio waveform sampled data, and to realize the conversion between the text data and musical data with high fidelity. In [7], the author presents a multi-domain wavelet network based on auto-encoder to realize the variety of musical styles in an unsupervised way. In [8], the author designs a self-supervised learning strategy to achieve the transfer of timbre.

The image processing ways consist of two main stages: the feature extraction of the original musical wave data and the construction of the transformation model. The common methods have Convolutional neural networks (CNN), long short-term memory networks (LSTM), and generative adversative networks (GAN). [9] proposes an encoder-decoder model with LSTM construction to realize the musical style transition, and uses the STFT algorithm to reconstruct the musical data. [10] presents a CNN model based on the continuous wavelet transform (CWT) to generate a new audio style from style audio. [11] proposes a GRU-GAN model to generate the chords music, and the use of the GRU method can realize the autonomously learning chords, finally, the experimental results show that the model has a good style presentation. [12] uses the constant Q transform (CQT) to extract the feature of musical timbre and utilizes the CycleGAN model to achieve the musical timbre transition from the orchestral instruments.

From the existing literature, the overall number of studies on the musical style transition is still relatively small, especially, the most representative of Chines Bamboo flute and flute.

Bamboo flute and flute are both well-known beautiful melodic instruments, but there are significant differences in structure and timbre between these two instruments. Capturing the complete feature of the music style data and improving the generalization ability of the model is the biggest challenge. To address these problems, we first propose a CycleGAN model for the style transition of the traditional Chinese Bamboo flute music, the overall design idea is shown in Figure 1 (the detailed model design is presented in Section 4). In

this model, we utilize STFT algorithm to realize the time-frequency transformation of the music waveform, and the spectrum image of the music waveform is better for extracting the features by the CNN model; in the musical waveform features conversion stage, we use the CycleGAN model to convert the music played by bamboo flutes into flute music, which better combines the timbre characteristics of bamboo flute and flute.

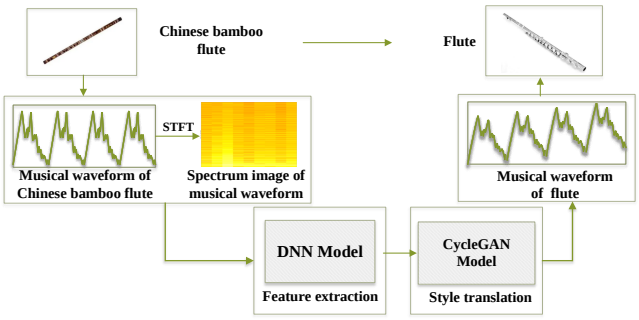

Fig. 1. The overall design idea of our proposed method.

The main contribution of this paper is as follows:

- It uses STFT algorithm to realize the time-frequency transformation of music waveform.
- It proposes a CNN model to extract the features of music waveform.
- It presents a CycleGAN method to convert the music played by bamboo flutes into flute music, and it is understood that this model is first proposed for the style transition of bamboo flute and flute.

II. Related Works

For the music style transition of bamboo flute and flute, the two most important things are the completeness of the feature extraction of the original music style and the construction of the transformation model. In this section, we mainly introduce our related works for our study.

A. Feature Extraction of Musical Style

There is a high time sampling rate in the waveform music data from the musical instruments (1600 samples per second), and it is time-dependent with the periodic, which is difficult to directly model in the time-domain musical waveform. In some literature, utilizing the abstract features representation of the waveform to indirectly modeling. [13] uses Mel spectrogram to map the features and to use a WaveNet model for the musical waveform synthesis. In [14], the Mel spectrogram is employed to realize the data preprocessing, and then to train the model. However, by Mel Scale, Spectrogram is dot multiplied with several Mel filters, which will add to the computational burden. STFT algorithm is another way to realize the feature extraction of musical data. In [15], the STFT algorithm is used to separate Logarithmic amplitude and instantaneous frequency from the frequency domain, and as the input of the GANs model, finally, to achieve efficient audio synthesis. In [16], the same method is employed for the rhythm recognition of Chinese musical instruments. However, the frequency-domain image can not represent the temporal coherence of musical style. We propose a CNN structure, which first uses the STFT algorithm to obtain the frequency-domain image of the musical waveform and use the CNN structure to extract the features of musical style in the bamboo flute.

B. Generative Adversarial Networks

GAN is a deep learning model that learns unsupervised on complex distributions. The Model has two modules: the Generative module (G) and the Discriminative module (D), and GAN obtains a good output by the mutual game learning of these two modules. The basic structure of GAN is shown in Figure 2. The generator generates fake data samples (images, audio, etc.) and tries to fool the discriminator. The discriminator tries to distinguish between the real and the fake samples. Both generators and discriminators are the neural networks, and the modules G and D compete to analyze, capture, and replicate the changes in the sampled data set. Repeat these steps, and in the process, the generator and discriminator get better at what they do with each repetition. In [17], GAN is used to realize the time features learning of music and to improve the stability of the synthetic multi-instrument music. In [18], GAN is employed to generate a dual track music generation model and a deep chord gated recurrent neuro generative adversarial neural for the music generation. In [19], GAN is used for music genre transfer and utilizing the single generator network to learn the many-to-many mappings of the different attribute domains. Though the GAN model has shown advantages in the field of music creation, there are some problems: the slow convergence and the poor model stability; G optimization needs enough gradient information; the training is difficult to converge.

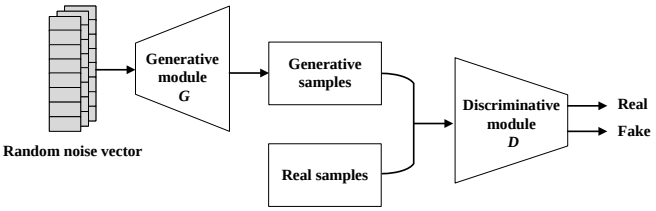

Fig. 2. The basic structure of GAN.

C. Learning of Musical Style Transfer

The research of musical style transfer in Chinese musical instruments acquires inspiration from image recognition. There are two methods: direct modeling by the musical waveform and indirect modeling by the abstract feature of the musical waveform. Direct modeling is technically challenging, so abstract feature learning and AI models are widely used in

the transformation model between the different styles of instruments. [20] presents an automatic music generation method, and it uses LSTM and GRU to improve the quality of the musical composition. [21] uses an enhanced LSTM model to capture the dynamics of the music rhythm. [22] proposes a music unsupervised style transfer method, and MFCC is used for recognizing the different characteristics of the musical styles. The GAN model has remarkable learning ability in unsupervised learning. [23] designs an LSTM-based GAN model to realize the transition melodies from lyrics, and the generator and discriminator use the deep LSTM structure, which improves the performance of the transition model. [24] presents a network units-based GAN method to achieve the melody music creation with emotional factors, and two discriminators are used for enhancing the GAN model. In [25], to improve the fidelity of music authors use a CycleGAN model for music genre transition. From our research, the learning of musical style transfer is currently relatively few, especially in the different types of instruments. In our paper, we first propose to use the CycleGAN model for the musical style transfer of traditional Chinese bamboo flute music.

III. Dataset

In this section, we mainly describe the musical sampled data set in detail and introduce the method of the data processing and storage respectively.

A. Data Acquisition

We gathered music that is played with flutes and music that is played with bamboo flutes using NetEase Cloud Music, which is the platform where music can be downloaded. We decide to use midi to represent audio data because it is relatively easy to be turned into arrays to be used as input for models.

B. Data Processing

After that, we used matrices to represent the data. We have to first sample the data with a sampling rate of sixteen-time steps per bar where a bar is a time segment that has a certain number of beats. This is because the music we obtained has different time signatures but it requires uniform time steps per bar to allow matrix representation.

For simplicity, we adjusted the velocity of all the music to the same value, assuring every note has the same loudness. This allows us to represent the notes easier by taking only on and off states of notes instead of specific values. We also merged all the tracks in the midi format into a single track that contains most of the original identity. To make training more simplistic, we normalized the pitch values and ignored the notes with pitches above C8 and below C1 since these pitches are rare. As a result, we have pitch values ranging between 0 and 1 with 84 possible values, and we got a matrix with size (16, 84) for every bar where 16 is the time steps and 84 is the pitch range.

Each training sample contains four bars and therefore forms a matrix of size (64, 84). To create these training samples, we first concatenated all the music data, which gave us a large matrix. Then we split this large matrix into smaller matrices of the target size and labeled them. After that, we divided the samples of each dataset into a training set and a testing set, with a ratio of approximately 4 to 1. In addition, we created a mixed dataset that contains all the samples from the two datasets for an alternate mode of training.

IV. Our Model

In this section, we mainly introduce the overall construction design of our proposed model, namely DNN-CycleGAN in detail. The function of DNN-CycleGAN is to complete the style features transfer of the bamboo flute to flute, which can generate a unique musical style with bamboo flute and flute. Our model framework includes: learning the representation of musical waveform style and style transfer model GAN-based DNN-CycleGAN.

A. Overall Framework Design of DNN-CycleGAN

The overall framework design of DNN-CycleGAN is presented in Figure 3. From Figure 3, there are four parts: data collection and storage, data pre-processing, the extraction of musical waveform style features, and the design of the CycleGAN model.

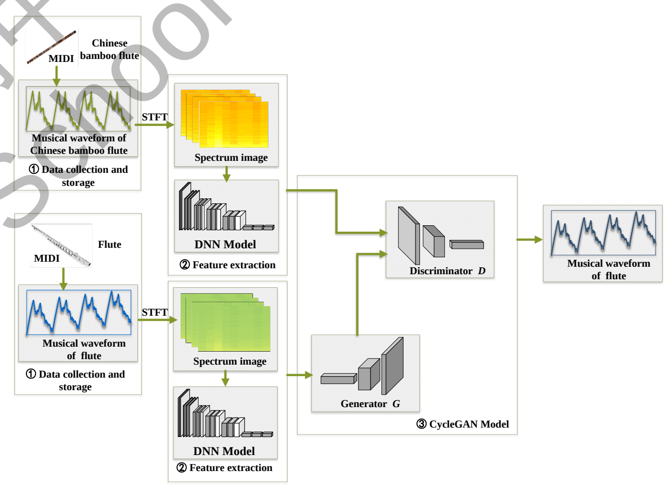

Fig. 3. The overall framework design of DNN-CycleGAN.

B. Representation Learning of Musical Waveform Style

In this paper, we design a DNN structure to complete the extraction of musical style features from Chinese bamboo flutes and flutes. This process is described as follows: firstly, the collected music waveform is converted into the spectrum image by utilizing the STFT algorithm; secondly, the spectrum image is the input data in the DNN model. As far as we know, the idea of STFT is that the long non-stationary stochastic process is regarded as the superposition of a series of short-time stochastic stationary signals, and the short-time characteristics can be achieved by adding a window function on time. In mathematics, STFT is described as follows:

$$G_Z(t,f) = \int_{-\infty}^{\infty} [z(u)g^*(u-t)]e^{j2\pi f(u)}du, \quad (1)$$

where $z(u)$ indicates the signal source and $g^*(u-t)$ indicates the time window function. Through STFT, the time-domain waveform of a music signal is transferred to a frequency-domain image, which will be the input data of the DNN model.

After that, we design a DNN model for extracting the musical waveform features, which takes its inspiration from the VGG-16 structure. The specific parameters of the model are as follows: the number of convolution layers, pooling layers, and fully-connected layers are 13, 5, and 3, respectively. Softmax is used as the activation function, and the size of the convolution kernel is 3×3. The detailed structural design is in Figure 4.

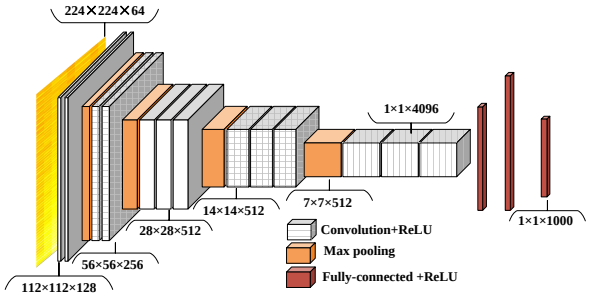

Fig. 4. The structural design of DNN.

Convolution layers In the DNN model, the function of the convolutional layer is to represent the features of musical style, and the key component is the convolution kernel. By the processing operation of the convolution kernel, the output of each convolution layer is as the new features mapping. It can be defined as the follows:

$$F(j)_i^n = f(\sum_{s \in N_{i(n-1)}} \sum_{(ku,kv) \in k^{(n)}} w_{js(ku,kv)}^s x_s^{(n-1)} (c+ku, r+kv) + Wb_i^{(n)}), \quad (2)$$

$$K^{(n)} = \{(ku,kv) \in M^2 | 0 < ku < K_w, 0 < kv < K_h\}, \quad (3)$$

where, the length and width of the convolution kernel are denoted by K_w and K_w, respectively. The size of the current network layer is n, and the offset of i^{th} features mapping in the convolution network layer is $Wb_i^{(n)}$. $s \in N_{i(n-1)}$ is the features mapping set in $n-1$ layer.

Max-pooling layer After the convolution operations, the max-pooling layer is used to reduce the dimension of the feature data by imitating the human vision system, which can effectively reduce network parameters and prevent overfitting. The max-pooling process is described as follows:

$$F(j)_i^n = MAX(F(j)_i^{(n-1)}). \quad (4)$$

Fully connected layer The fully connected layers are after the Convolution layers and the Max-pooling layers. The fully connected layers can integrate the local information with the category distinction in the convolution layer and pooling layer, then improve the DNN network performance. ReLU is used as the activation function, this process is described as follows:

$$o_i^n = f(\sum_{j=1}^{M} w_{ij}^{(n)} b_i^{(n)}). \quad (5)$$

C. The Framework Design of CycleGAN

In this paper, we use CycleGAN to complete the style transfer in Chinese bamboo flute music and flute music, which can solve the problem of the music style transfer of the unmatched data. The aim of CycleGAN is to learn the data conversion functions $F(x)$ and $G(y)$ between two different types of domains X and Y. $F(x)$ is used to the sampled data $x \in X$ convert into the elements of Y, $F(x) : x \to y$. $G(x)$ is used to the sampled data $y \in Y$ convert into the elements of X, $G(x) : y \to x$. CycleGAN is a ring structure and is mainly composed of two generators G and two discriminators D. The details of the model structure are presented in Figure 5.

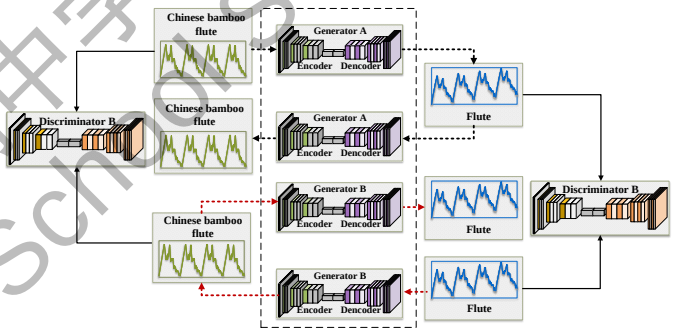

Fig. 5. The details structure of CycleGAN.

Generator: the generator consists of an encoder, a converter, and a decoder. Encoding: the first step, the convolutional neural network is used to extract features from the input image. The image is compressed into 256 64*64 feature vectors. Converter: the feature vectors of the image in the DA domain are converted to the feature vectors in the DB domain by combining the non-similar features of the image. We use six layers of the Reset module, each of which is a neural network layer composed of two convolutional layers, which can achieve the goal of preserving the original image features during transformation. Decoding: Using the deconvolution layer to restore the low-level features from the feature vectors, and finally get the generated image. The structure of the generator is shown in Figure 6.

Discriminator: the discriminator takes an image as input and tries to predict whether it is the original image or the output image of the generator. The discriminator itself belongs to a convolutional network, which needs to extract features from images and then determine whether the extracted features belong to a specific category by adding a convolutional layer

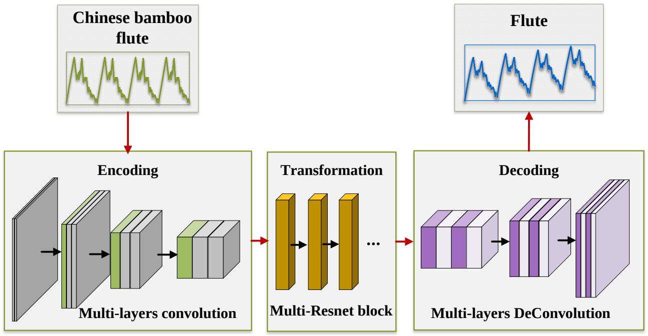

Fig. 6. The structure of generator.

that produces one-dimensional output. The structure of the discriminator is shown in Figure 7.

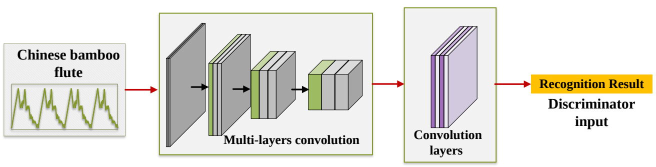

Fig. 7. The structure of discriminator.

Loss: In this paper, we use the DNN-CycleGAN model to achieve music style transfer. The core is the cooperation of two GAN networks. The generator is responsible for generating a spurious sample, $G_{x\to y}$, and the discriminator attempts to distinguish the difference between the generated sample $G_{x\to y}$ and the actual sample x. In the process of music style transfer, to promote the preservation of effective audio information, CycleGAN supplements the loss of identity mapping. The specific expressions are as follows:

Adversarial Loss: to make the transfer music features indistinguishable from the original target, the adversarial losses are used as follows:

$$\begin{aligned} Loss_{adv} = (G_{X\to Y}, D_Y) = E_{y\sim P_{Y(y)}[logD_{Y(y)}]} \\ +Ex \sim P_{X(x)}[log(1 - D_Y(G_{X\to Y(x)}))], \end{aligned} \tag{6}$$

where the discriminator D_Y seeks the true music style feature by maximizing the adversarial loss and making the best decision boundary between the features and transformation features. By minimizing adversarial losses, the generator $G_{X\to Y}$ generates the features to cheat D_Y.

Cycle-consistency Loss: to regularize the mapping, Cycle-consistency loss is used. It uses the forward and inverse correlation of dual mapping to improve the consistency loss of the model. The cycle-consistency loss helps the generators $G_{X\to Y}$ and $G_{Y\to X}$ find the best pairing for the (X, Y) combination in the form of a transformation of the $X \to Y$ cycle.

$$\begin{aligned} Loss_{cyc}(G_{X\to Y}, G_{Y\to X}) = \\ E_{x\sim P_X}(x)[||G_{Y\to X}(G_{X\to Y}(x)) - x||_1] \\ +Ey \sim P_Y(y)[||GX \to Y(GY \to X(y)) - y||_1], \end{aligned} \tag{7}$$

Identitymapping Loss: to further preserve the input, use identity mapping loss:

$$\begin{aligned} Loss_{id}(G_{X\to Y}, G_{Y\to X}) = E_{x\sim P_X}(x)[||G_{Y\to X}(x) - x||_1] \\ +Ey \sim P_Y(y)[||GX \to Y - y||_1], \end{aligned} \tag{8}$$

Therefore, the total loss can be written as a linear combination of the above three losses:

$$\begin{aligned} Loss_{full} = Loss_{adv}(G_{X\to Y}, D_Y) \\ +Loss_{adv}(G_{Y\to X}, D_X) \\ +\lambda_{cyc}Lcyc(G_{X\to Y}, G_{Y\to X}) \\ +\lambda_{id}Lid(G_{X\to Y}, G_{Y\to X}), \end{aligned} \tag{9}$$

where, $Lfull$indicates the final loss; λ_{cyc} and λ_{id} is a hyper-parameter, and to control the significance of the related loss.

V. Experiments

In this section, we focus on the presentation and discussion of experimental results, the model training and test process, and performance evaluation results.

A. *Model Training*

During the training process, we used the Adam optimizer with an initial learning rate of 0.0002 and momentum decay rates of 0.5 and 0.999. The batch size in all training processes is set to 16. The coefficient used to weigh the cycle consistency loss, which is often expressed as lambda, equals 10 in our model. The coefficient used to weight the extra discriminator loss, which is often expressed as gamma, is 1. We did five pieces of training in various settings. The first model is trained with a model that has two generators and two discriminators described in section 3. We trained this model with 8 epochs. The second model is trained with an identical structure as the first but with 10 epochs. The third model is trained with the extra discriminators and the rest of the architecture of the model is the same as the first two models. This model is trained with 10 epochs. The fourth model is trained with the extra discriminators but used Softmax as the activation function in discriminators. This model is trained with 10 epochs. The fifth model is trained with extra discriminators and LeakyRelu activation function for discriminators, but with only two convolutional layers instead of three. This model is trained with 10 epochs.

B. Results and Discussion

We choose a classic Chinese bamboo flute music, $JasmineFlower$ as the experimental data to verify our method. The time-domain waveform of Chinese bamboo flute music is shown in Figure 8. After that, the time-domain waveform of the music style is inputted into the DNN-CycleGAN for the musical style transfer. The transformed music style waveform is shown in Figure 9.

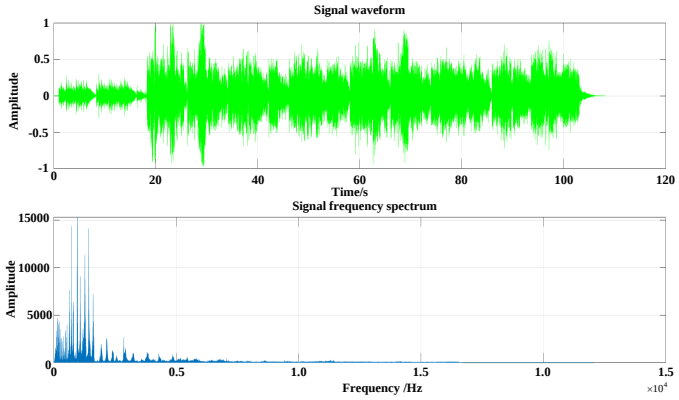

Fig. 8. Chinese bamboo flute music waveform.

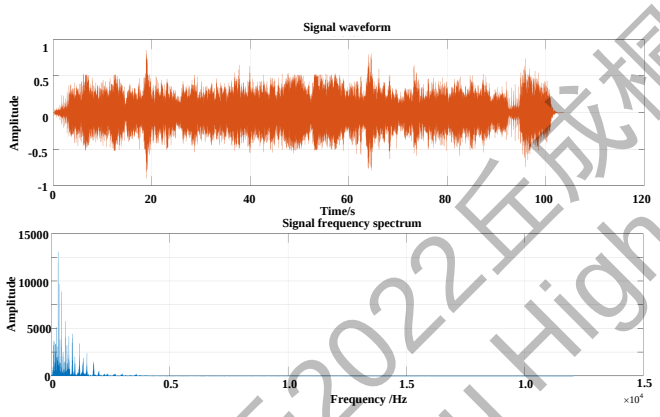

Fig. 9. The transformed music style waveform.

Overall, the loss of the generator is substantially higher than the loss of the discriminator. However, we tried various ways to improve the situation and some of them worked well since the difference in the two losses is reduced effectively. We plotted the loss of generators and discriminators while training, the results are shown in Figures 10, 11, 12, 13, and 14. The first model (no extra discriminators, 8 epochs), second model (no extra discriminators, 10 epochs), third model (extra discriminators, 10 epochs), fourth model (extra discriminators, Softmax activation, 10 epochs), and fifth model (extra discriminators, 2-layer discriminator, 10 epochs).

From the first two results, we see that increasing the number of epochs did not decrease the loss of the discriminators or the loss of generators and it did not solve the problem where the generator and the discriminator are unbalanced.

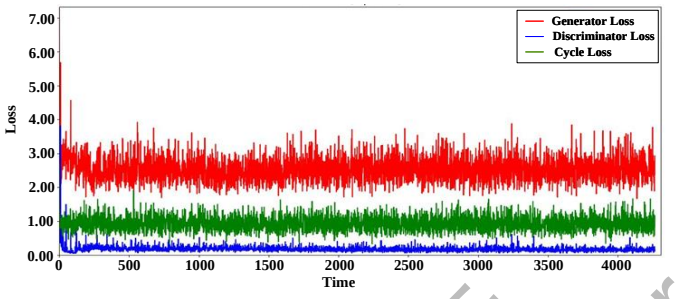

Fig. 10. First model (no extra discriminators, 8 epochs).

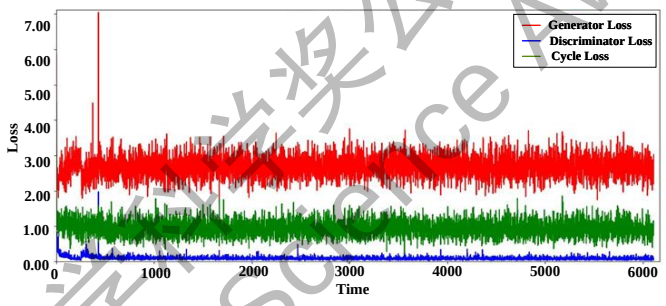

Fig. 11. Second model (no extra discriminators, 10 epochs).

Comparing this model with the second, we see that the difference in the losses is reduced and the generator loss decreased by a large amount. Then we used the third model (extra discriminators, 3-layer discriminator, Leaky Relu activation, 10 epochs) to test performance on the testing set and generated samples. We found that most of the styles of the flute are successfully transferred to the bamboo flute and vice versa.

Comparing this model with the third, we see that the generator loss is higher, so we conclude that Softmax activation is less suitable than LeakyRelu activation in this model. Comparing this model with the third one, we see that the difference in the losses has decreased but the generator loss has not, so reducing the layers of the discriminator did not work well in this situation.

The

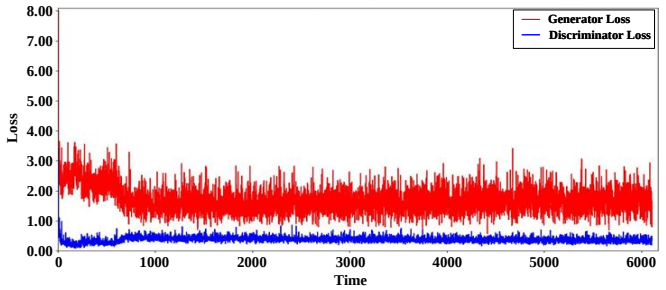

Fig. 12. Third model (extra discriminators, 10 epochs)

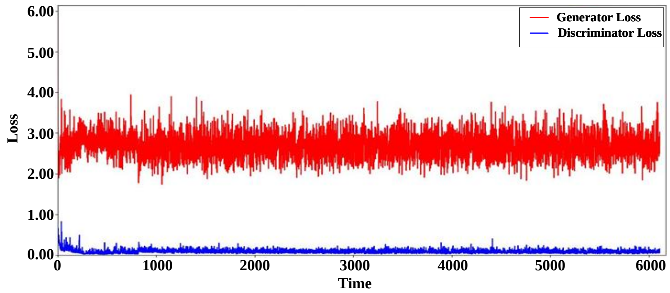

Fig. 13. Fourth model (extra discriminators, Softmax activation, 10 epochs).

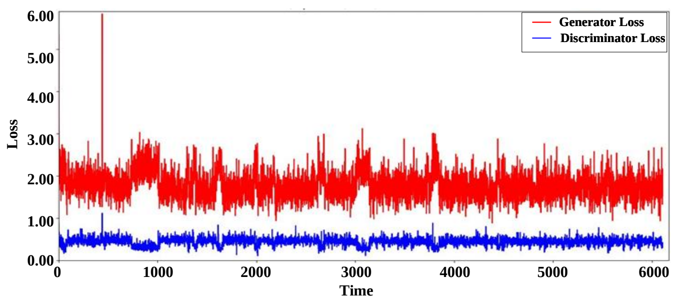

Fig. 14. Fifth model (extra discriminators, 2-layer discriminator, 10 epochs).

C. Model Subjective Evaluation

To better verify the performance of our model, we adopts the subjective evaluation to evaluate the performance of the model. We employ MOS as the subjective evaluation criterion.

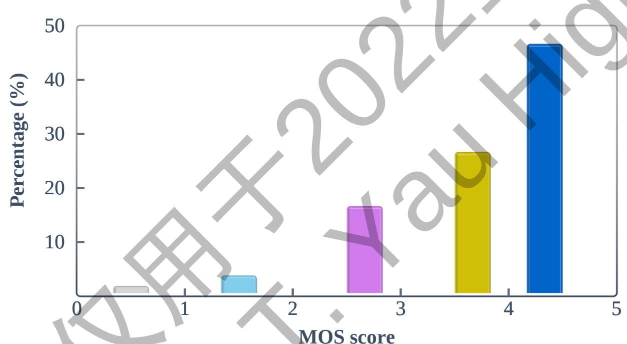

Fig. 15. The percentage of subjective evaluation.

The specific scoring rule is set as follows: making subjective judgments about the original music and the music generated by the model in different people. The score $x \in [1, 2, 3, 4, 5]$, the higher the score, the better the effect. The experimental test results are shown in Figure 15. From Figure 8, the scores between 4 and 5 account for 48% of the total score, and the scores between 3 and 4 account for 38% of the total score, which illustrates our model has good performance.

VI. Conclusion

In this paper, we propose a machine learning method DNN-CycleGAN to transfer music from the genre that consists of the bamboo flute to the genre that consists of the flute. The model includes a music style feature learning module based on DNN and a CycleGAN for the transferring of music style. A considerable amount of experiments have approved that transferring music from a Chinese genre to a western genre can be done successfully. In the future, we can transfer music from other domains and genres or add velocities into account instead of turning the velocities of all notes the same as we did this time.

References

[1] Yan L., Zheng W.B., Gou C., and Wang F.Y., “IPGAN: Identity-Preservation Generative Adversarial Network for unsupervised photo-to-caricature translation,” KNOWLEDGE-BASED SYSTEMS, vol. 24, 108223, 2022.

[2] Mehri S., Kumar K., Gulrajani I. , Kumar R., Jain S., Sotelo J., Courville A., and Bengio Y., “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model,” KNOWLEDGE-BASED SYSTEMS, 2016.

[3] Mital P. K., “Applying visual domain style transfer and texture synthesis techniques to audio: insights and challenges,” NEURAL COMPUTING & APPLICATIONS, vol. 32, pp. 1051-1065, 2017.

[4] Liu X.H., Delany S.J., and McKeever S, “Sound Transformation: Applying Image Neural Style Transfer Networks to Audio Spectograms,” NEURAL COMPUTING & APPLICATIONS, vol. 11679 , pp. 330-341, 2019.

[5] Fatima S. M., Shehzad M., Murtuza S. S., and Raza S. S., “Neural Style Transfer Based Voice Mimicking for Personalized Audio Stories,” the processing paper at AI4TV ’20: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery, pp. 11–16, 2020.

[6] Oord Avd, Dieleman S., Zen H., Simonyan K., and Kavukcuoglu K., “WaveNet: A Generative Model for Raw Audio,” 2016.

[7] Mor N., Wolf L., Polyak A., and Taigman Y., “A Universal Music Translation Network,” 2018.

[8] Cifka O, Ozerov A, Simsekli U, Richard G , “SELF-SUPERVISED VQ-VAE FOR ONE-SHOT MUSIC STYLE TRANSFER,” the processing paper at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 96–100, JUN 2021.

[9] Modrzejewski M., Bereda, K. ,and Rokita P., “Efficient Recurrent Neural Network Architecture for Musical Style Transfer,” the processing paper at 20th International Conference on Artificial Intelligence and Soft Computing (ICAISC), vol. 12854, pp. 124–132, JUN 2021.

[10] Chen J., Yang, G., Zhao H., and Ramasamy M., “Audio style transfer using shallow convolutional networks and random filters,” Multimedia Tools and Applications, vol. 79, pp. 15043–15057, JUN 2020.

[11] Li XR and Niu YZ, “Research on Chord-Constrained Two-Track Music Generation Based on Improved GAN Networks,” SCIENTIFIC PROGRAMMING, vol. 2022, MAR 2022.

[12] Shen J., Pang R., Weiss Ron J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhang Y., Wang Y., and Skerry-Ryan R., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” the processing paper at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783, MAR 2017.

[13] Verma P. and Smith Julius O, “Neural Style Transfer for Audio Spectograms,” 2017.

[14] Huang S., Li Q., Anil, C., Bao, X., and Grosse R. B., “TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer,”, 2018.

[15] Engel J., Agrawal K. K., Chen S., Gulrajani I., Donahue C., and Roberts, A., “GANSynth: Adversarial Neural Audio Synthesis,”, 2019.

[16] Lin S., Wu W., and Xie, L., “Automatic Identification of Chinese Musical Instruments,”, the processing paper at 2nd International Symposium on Intelligence Computation and Application (ISICA 2007), pp. 402–404, 2007.

[17] Guan F., Yu, C., and Yang, S. A GAN Model With Self-attention Mechanism To Generate Multi-instruments Symbolic Music, the processing paper at International Joint Conference on Neural Networks (IJCNN), Budapest, HUNGARY, 2019.
[18] Hu JX, Ge ZH ,and Wang XH, "The Psychological Education Strategy of Music Generation and Creation by Generative Confrontation Network under Deep Learning," the processing paper at 19th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT), vol. 2022, 3847415, 2022.
[19] Pezzat M., Perez-Meana H., Nakashika T., and Nakano M., "Many-to-Many Symbolic Multi-Track Music Genre Transfer," COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE, vol. 327, pp. 272–281, 2020.
[20] Gunawan Aas, Iman A. P., and Suhartono, D., "Automatic Music Generator Using Recurrent Neural Network," International Journal of Computational Intelligence Systems, vol. 13, 1, pp. 645, 2020.
[21] Jedrzejewska M. K., Zjawinski, A., and Stasiak B., "Generating Musical Expression of MIDI Music with LSTM Neural Network," the processing paper at 11th International Conference on Human System Interaction (HSI), Gdansk Univ Technol, Gdansk, POLAND, pp. 132–138, JUL 2018.
[22] Lu C. Y., Xue M. X., Chang C. C., Lee C. R. ,and Su, L., "Play as You Like: Timbre-enhanced Multi-modal Music Style Transfer," the processing paper at 33rd AAAI Conference on Artificial Intelligence / 31st Innovative Applications of Artificial Intelligence Conference / 9th AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, pp. 1061–1068, JAN 2019.
[23] Yu Y., Srivastava A. ,and Canales, S., "Conditional LSTM-GAN for Melody Generation from Lyrics," ACM TRANSACTIONS ON MULTIMEDIA COMPUTING COMMUNICATIONS AND APPLICATIONS, vol. 17, 35, 2021.
[24] Li S., Jang S., Sung, Y., "Conditional LSTM-GAN for Melody Generation from Lyrics," Automatic Melody Composition Using Enhanced GAN, vol. 7, 10, 2019.
[25] Brunner G., Wang, Y., Wattenhofer, R., and Zhao, S. "Symbolic Music Genre Transfer with CycleGAN," the processing paper at 30th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Volos, GREECE, pp. 786–793, 2018.

VII. Contribution

- Ziqi Qin: Conceptualization, Methodology, Software, Data curation, Writing- Original draft, Formal analysis.
- Yixuan Huang: Software, Validation, Formal analysis, Resources.
- Zijie Zheng: Writing- Reviewing and Editing, Supervision,Visualization.