*Journal of Network Intelligence*
**Taiwan Ubiquitous Information**
 ISSN 2414-8105
Volume **10**, Number **2**, May **2025**

# Digital Music Score Detection Based on Attention Mechanism Generating Adversarial Networks

Jin-Hua Li

School of Music
Sichuan University of Science and Engineering, Zigong 643000, P. R. China
18782138837@163.com

Xue-Jiao Luo*

School of Music
Sichuan University of Science and Engineering, Zigong 643000, P. R. China
15828677347@163.com

Qian Wang

Department of Music and Performing Arts
Philippines Sao Paulo University, Manila 1004, Philippines
m18381357773_1@163.com

*Corresponding author: Xue-Jiao Luo
Received April 24, 2024, revised August 25, 2024, accepted December 2, 2024.

Abstract. *Music score detection is very important to promote the intelligentization of music. Traditional music score detection methods have the issue of imbalance of note types, which leads to poor model recognition effect. The emergence of Generative Adversarial Network (GAN) can provide a novel solution to the data imbalance, but the original GAN have the issues of unstable training and insufficient feature extraction when generating data. Thus, this article firstly adopts Wasserstein distance instead of JS dispersion, which is a measure of data distribution in GAN, to enhance the training process of the network; and introduces the fully connected module instead of the basic module of the generator, which optimizes the feature capturing ability of the network. Next, preprocessing the music score image to reduce and remove irrelevant features. The squeezing and excitation modules are appended to the generator of the improved GAN to explicitly model the interdependence between the feature channels, and then the discriminator incorporates a Bidirectional Gating Unit (BiGRU), to capture the key features of the input note sequences, which are adopted as inputs by fusing them with the attentional output features. At last, the output fused characteristics are linearized using a fully connected layer to output the final binary classification detection results. The experimental outcome indicates that the suggested method has low note error rate, sequence error rate, training time consuming and network loss, which proves the efficiency and accuracy of the designed method.*
**Keywords:** Music score detection; Generative adversarial network; Wasserstein distance; Attention mechanism; Bidirectional gating unit

1. **Introduction.** Music is a form of natural language that expresses human emotions and reflects the spiritual culture of human production and life, and sheet music is the material carrier of this spiritual content [1]. Sheet music is divided into pentatonic score and simple score, the international common way to record music is pentatonic score, and in China, simple score is the most commonly used notation besides pentatonic score, and

most of the musical works from ancient times to the present day have been passed down in the form of paper simple score, which has been widely promoted and preserved. In the early days, the digitization of music scores was mainly done by manual input, which was demanding for the recorders, complicated and inefficient, and the slowness of manual input contradicted with the fast information processing of computers [2, 3, 4]. To address this issue, computerized music score detection technology has emerged to replace manual input. The music score detection is dedicated to the research of allowing computers to directly read music symbols in image files [5], and to segment and recognize the music score symbols by computers and save them in MusicXML or MIDI format, which is of great value in the fields of digital music libraries [6], music information retrieval [7], and music search engines [8], and so on.

1.1. **Related work.** With the deepening of deep learning research, scholars have applied deep learning methods to various stages of music score detection to optimize the effect. Li and Zheng [9] designed a selective self-encoder learning binarization conversion for simple spectrum images, which outperforms the traditional binarization methods, but is prone to errors in the foreground pixel edges. Sarkar et al. [10] used convolutional neural networks to recognize handwritten simple symbols. Calvo-Zaragoza et al. [11] used a combination of Hidden Markov Models and Artificial Neural Networks to detect handwritten sketches, but the effectiveness was limited by the pre-segmentation process of the notes. Adavanne et al. [12] used multiscale residual Convolutional Neural Networks (CNN) to extract the note characteristics of the simple notation images, and utilized the two-way simplex recurrent unit network to recognize the note features, which accelerates the training's convergence rate. Bisharad and Laskar [13] adopted the convolutional Recurrent Neural Network (RNN) to detect the note features in simple notation images, which speeds up the training rate. Han et al. [14] combined a residual recurrent CNN block with a recurrent encoder-decoder network to detect musical notation sequences. Huang et al. [15] suggested a deep neural network-based method to detect the music score notes, and achieved the recognition of the short score in the short score image by using a two-way simple RNN. Cortes et al. [16] proposed a stable path-based approach for detecting spectral lines in a music score, which has some limitations in the skew correction of a music score image. Rajesh and Nalini [17] sliced the music score in simple scores, and combined it with recurrent neural networks to identify the pitch and timing of the notes in the sequence. Lee and Le [18] used ResNet-101 as a backbone network to capture the characteristics of the music score image, and then performed regression and detection based on the feature maps, but the accuracy of the detection is not high. To enhance the detection efficiency of music scores, Tomaz Neves et al. [19] offered a Generative Adversarial Network (GAN) based method for music score detection under small sample conditions and fused CNN to extract short spectrum features. Wang et al. [20] used PCA technique to reduce the dimensionality of the music score image and then used GAN to detect the notes, but the detection effect is not good. Nakamura et al. [21] used CNN instead of multilayer perceptron in GAN to improve the efficiency of music score recognition. Lattner and Nistal [22] suggested to use a combination of statistical recognition and note features recognized by GAN structure as a classification basis for music score recognition. Marafioti et al. [23] used a model that combines a GAN with a connected time-series classifier to achieve note recognition.

1.2. **Contribution.** The above mentioned deep learning based music score detection method has the issue of imbalance of note types, which results in poor detection effect. When training a deep learning model, if the dataset used is not uniform in the distribution of note types, i.e., some types of notes occur much more frequently than others, it can

result in the model recognizing common notes better and uncommon notes worse. Aiming at the above issues, this article suggests a digital music score detection method based on attention mechanism GAN. Firstly, the Wasserstein distance is adopted as an index to measure the distance between the generated dispersion and the actual dispersion, and the artifacts are eliminated by introducing the fully connected module instead of the basic module of the generator. Secondly, gray scale conversion, sharpening and binarization are carried out on the music score image to reduce irrelevant features. Next, squeezing and excitation modules are appended to the GAN's generator to explicitly model the interdependence between the feature channels, and then the discriminator incorporates the BiGRU to obtain the key features of the input picture, and by fusing the attentional output characteristics with the key features as inputs to the subsequent model. Finally, the fused features output from the model are linearly operated using a fully connected layer to output the final binary classification detection results.

## 2. **Theoretical analysis.**

### 2.1. **Generating adversarial network.**

Two network models that are continuously trained during the confrontation process constitute a generative adversarial network [24], where the network used to generate sample data is called generator $G$, which learns the distribution of the original dataset to fit and generate a completely new distribution, and the other network identifies between the actual and the generated data, which is denoted as discriminator $D$. The basic framework of the GAN is indicated in Figure 1.

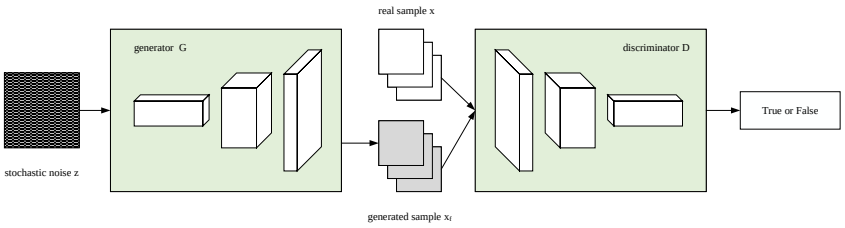


Figure 1. The network framework of GAN.

Let the random noise input to the generator be $a$, and the prior distribution of the noise be $p_a$, which conforms to a normal or mean distribution. The generator is denoted by $G(a;\delta_g)$, and the discriminator is denoted by $D(x;\delta_d)$, where $\delta_g$ is the parameter of the $G$ and $\delta_d$ is the parameter of the $D$. Assuming that the final data sample dispersion generated by the $G$ is $p_g$, and the real data sample distribution is $p_{rdata}$, the generator is to let $p_g$ achieve the purpose of fitting $p_{rdata}$ as much as possible. Generate the objective function for adversarial network optimization as bellow.

$$\min_G \max_D J(D,G) = E_{x\sim p_{rdata}}\left[\log D(x)\right] + E_{a\sim p_a}\left[\log(1-D(G(a)))\right] \tag{1}$$

where $E_{x\sim p_{rdata}}\left[\log D(x)\right]$ is the $D$'s discriminant expectation of the real data sample, $E_{a\sim p_a}\left[\log(1-D(G(a)))\right]$ is the $D$'s discriminant expectation of the generated data sample, and $J(D,G)$ is the gap between the two distributions judged by the discriminator. When the data distribution is a continuous probability distribution, through the relationship between the probability density function and the expectation, the above equation can be changed into the Equation (2).

$$V(G,D) = \int_x p_{rdata}(x)\cdot\log(D(x)) + p_g(x)\cdot\log(1-D(x))\,dx \tag{2}$$

During training process, the $G$ and the $D$ are trained alternately, and when the discriminator can correctly distinguish between true and false data samples, $V(G, D)$ takes the extreme value. The solution of the discriminator is as follows.

$$D_G^*(x) = \frac{p_{rdata}(x)}{p_{rdata}(x) + p_g(x)} \tag{3}$$

2.2. **Musical score.** The musical score consists of seven Arabic numerals to indicate the pitch of the notes and the interrelationship between the notes [24], i.e. 1, 2, 3, 4, 5, 6, 7. However, only seven notes, do, re, mi, fa, sol, la, si, cannot constitute a sheet of music, and if there are no pitches or basses in the whole score, the music will lose its soul and become tasteless. Therefore, there are some high and low tones in the actual music score, that is, in the simple score, you will find a '-' above the note, which means that the tone is raised by one octave, and a '-' below the basic note, which means that the tone is lowered by one octave, and at the same time, there are scales in the simple score. There are many types of scales, but most scales are the basic natural major scale. This is indicated in Figure 2.

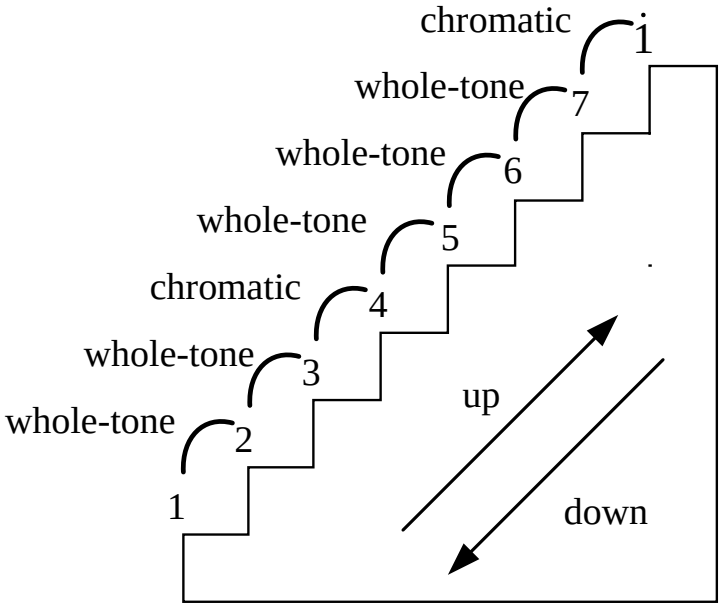


Figure 2. Sample diagram of a music score

3. **Optimization for generating adversarial network.** Focusing on the issues of the traditional GAN, such as difficult training process, smooth edges of reconstructed images and distortion of details, the Wasserstein distance is used to replace the JS dispersion. To optimize the training procedure, combining with dense connectivity network's idea, the fully connected module is introduced to replace the basic module of the generator, so as to improve the feature capturing the network's capability.

The Wasserstein distance, which is known as the Earth-Mover distance [25], measures the minimal cost needed to "transport" mass from one distribution to another as defined below.

$$W(p_r, p_g) = \inf_{\lambda \sim \Pi(p_r, p_g)} E(x, y) \sim \lambda[\|x - y\|] \tag{4}$$

where $\Pi(p_r, p_g)$ is each probable joint dispersion's set when $p_r$ and $p_g$ are combined. For every probable joint dispersion $\lambda$, the actual instance $x$ and the generated instance $y$ is able

to be instanced from $(x, y) \sim \lambda$. The distance between them is $\|x-y\|$, and the expectation of the samples over the distance for the jointed dispersion is $E(x, y) \sim \lambda[\|x - y\|]$.

For the purpose of adapting the GAN network, a simple transformation of the Wasserstein distance is performed as indicated in Equation (5).

$$W(p_r, p_g) = \frac{1}{L} \sup_{\|f\|_K \leq L} E_{x\sim p_r}[f(x)] - E_{x\sim p_g}[f(x)] \tag{5}$$

where $L$ is the Lipschitz constant. If the Lipschitz constant $\|f\|_K$ of a function $f$ is less than $L$, then for each possible $f$ meeting the status, take an upper bound on $E_{x\sim p_r}[f(x)]-E_{x\sim p_g}[f(x)]$ and allocate by $L$. The function $f_v$ is then defined with a set of parameters $v$ as follows.

$$L \cdot W(p_r, p_g) \approx \max_{v:\|f_v\|_K} E_{x\sim p_r}[f_v(x)] - E_{x\sim p_g}[f_v(x)] \tag{6}$$

Then the Wasserstein distance is adopted to optimize the loss operation to avoid the discriminative network evaluating the image generated by the generator network only as a separate classification problem. The optimized GAN discriminative model feeds the image GF formed by the generator and the actual image TF into the discriminator, extracts the features through the first several layers of convolution, and then feeds the extracted features into the later convolutional layers for dimensionality reduction, removes the sigmoid level, and uses the fully-connected level to directly output the discriminative process's final results.

The loss operation of the improved GAN makes up of two parts, the content loss and the adversarial loss, which are weighted with certain weights during the training process.

$$L^{GF} = L_X^{GF} + 10^{-3} L_g^{GF} \tag{7}$$

where the first term $L_X^{GF}$ is the content loss.

The second term is generating the adversarial loss, which is to generate data distributions that make it difficult for the discriminator to distinguish truth from falsehood, and is defined as follows.

$$L_G^{GF} = \sum_{n=1}^{N} -\log D_{\delta_G}\left(G_{\delta_G}\left(I^{TF}\right)\right) \tag{8}$$

In a traditional GAN, the discriminator is a binary classifier that decides whether the input instances are actual or generated. However, in an optimized GAN, the discriminator's goal is to approximate the Wasserstein distance, which is a regression task, rather than minimizing the cross-entropy loss function as in a traditional GAN. The loss functions of the generative and adversarial networks are defined according to the Wasserstein distance as follows.

$$L_G = -E_{x\sim p_g}\left[f_v(x)\right] \tag{9}$$

$$L_D = E_{x\sim p_r}\left[f_v(x)\right] - E_{x\sim p_g}\left[f_v(x)\right] \tag{10}$$

## 4. **Research on music sketch recognition method based on improved GAN.**

4.1. **Pre-processing of music score image.** Intending to the issue of small scale and poor diversity of traditional music score recognition methods, this article designs a digital music score detection method based on attention mechanism GAN. Firstly, preprocessing the music score image, and then adding squeezing and excitation module in the generator of GAN to extract features from the preprocessed image. In order to enhance the features extracted by the generator, the discriminator incorporates the BiGRU to obtain the original features of the input note sequence, and finally adopts the fully connected level to perform linear operations on the fused features output from the model to output the final

binary classification detectiontion results. The entire model of the suggested method is indicated in Figure 3.

The main task of preprocessing is to reduce and remove irrelevant features in the recognition process of a sketchy image, specifically, it is divided into three steps: grayscale conversion, sharpening, and binarization.

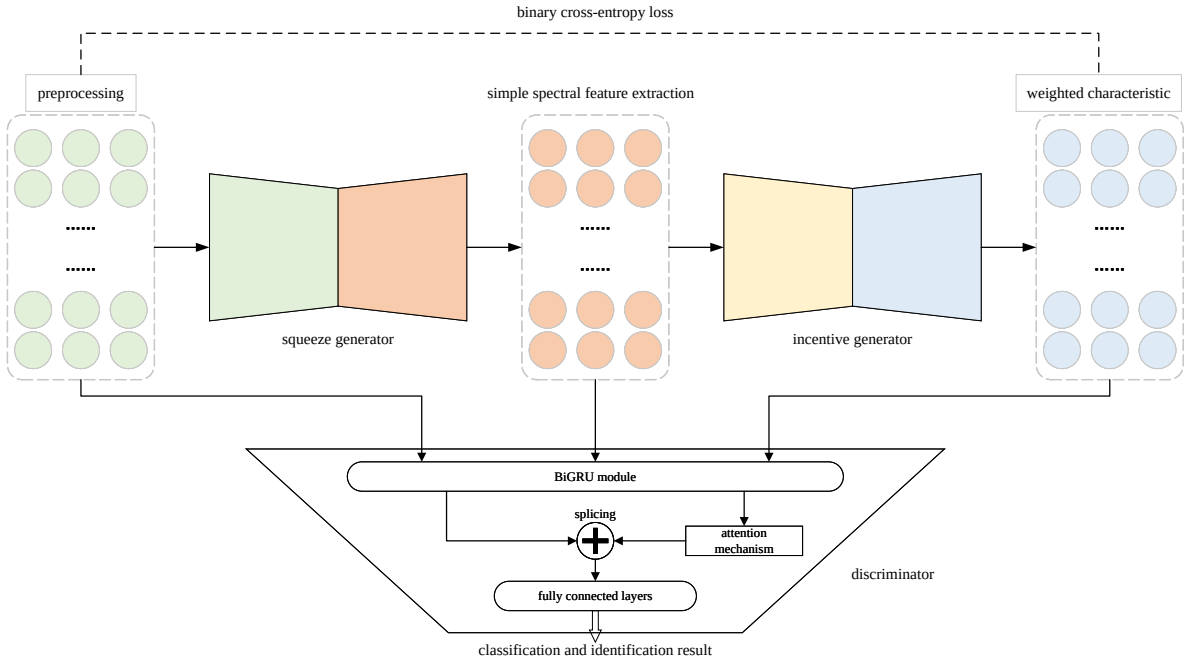


Figure 3. The whole model of the designed method

(1) Gray scale conversion. The information obtained by the camera is mostly 32-bit true color information, but the music score is only in black and white. Before starting the detection of the music score, use the weighted average method to convert the color image into 8-bit grayscale, which converts the RGB three-dimensional information of the image into one-dimensional, greatly reducing the amount of computation, and improving the efficiency of the symbol detection in the later stage.

$$y = (0.299)R + (0.587)G + (0.114)B \tag{11}$$

(2) Sharpening. The image of a music score is noisy and cannot clearly highlight the features of the spectrum. In this paper, the Laplace operator [26] is used to sharpen the image to highlight the details of the image, to enhance the regions of the image with sudden changes in grayscale, and to attenuate the regions with slow changes in grayscale. The Laplace transform of a 2D image is defined as Equation (12), which is the sum of the second order derivatives of the image in the $x, y$ direction.

$$Laplace(f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{12}$$

(3) Binarization. To further reduce the amount of computation, the background of the score is set to black (with a value of 0) and the notes are set to white (with a value of 255 in the 8-bit image), which is convenient for the subsequent feature extraction and image recognition. The binarization belongs to the global threshold segmentation method, which selects the threshold value $T$ and the extreme values $a$ and $b$ for the whole image, and for any pixel $f(i, j)$ at coordinate $(i, j)$ in the music score image, if the pixel value at

coordinate $(i, j)$ in the music score image is $g(i, j)$, the following equation is used.

$$\begin{cases} g(i,j) = a, & f(i,j) < T \\ g(i,j) = b, & f(i,j) \geq T \end{cases} \tag{13}$$

4.2. **GAN-based music score feature extraction.** After preprocessing the music score image, the features of the image need to be captured for image description. However, for a music score image, especially a low-quality music score image, spectral lines, non-note ink blots, noise, etc. belong to the interference information, so the extraction of interference information such as spectral lines and noise should be weakened as much as possible when performing the feature extraction, so as to enhance the feature representation capability of the model.

To further improve the detection performance, GAN incorporates a squeezing and excitation module in the generator [27], which explicitly models the interdependence between feature channels, and learns to promote useful features and suppress those that are not very useful for the task at hand. Suppose $X(x_1, x_2, ..., x_d)$ is the characteristic map of the input spectral image, with height $h$, width $v$, and the amount of characteristic channels $d$. $F_{tq}$ is performed on $X$, and a feature description vector $Y(y_1, y_2, ..., y_d)$ with the global information of the image is obtained through the global pooling layer, whose size is changed to $1 \times 1$, and the number of feature channels is still $d$. The equation for $F_{tq}$ operation is as follows.

$$Y_d = F_{tq}(x_d) = \frac{1}{V \times H} \sum_{i=1}^{V} \sum_{j=1}^{H} x_d(i, j) \tag{14}$$

where $Y_d$ is the characteristic vector of the $d$-th channel of the characteristic description vector $Y$ and $x_d$ is the characteristic map of the $d$-th channel of the characteristic map $X$.

Then the excitation operation $F_{ex}$ is performed on $Y_d$, i.e., the new $1 \times 1 \times d$'s feature description vector $T(t_1, t_2, ..., t_d)$ with information about the importance of each channel of the image is obtained by learning the weight matrix $V_1, V_2$ through the two fully connected layers. The equation for $F_{ex}$ is as follows.

$$T = F_{ex}(Y, V) = \delta\big(g(Y, V)\big) = \delta\big(V_2 \sigma(V_1, Y)\big) \tag{15}$$

where $\sigma$ is the ReLU activation operation, $\delta$ is the Tanh activation operation, $V_1 \in R^{(d/r) \times d}$ is the dimension reduction parameter, $V_2 \in R^{d \times (d/r)}$ is the dimension increase parameter, and $r$ is the reduction factor.

A multiplication operation in $F_{scale}$, channel dimension is performed on the squeezed excitation processed feature description vector $T$ and the original feature map $X$ to obtain a weighted feature map $\tilde{X}$.

$$\tilde{x}_d = F_{scale}(x_d, t_d) = t_d x_d \tag{16}$$

4.3. **GAN-based music score image classification and detection.** To enhance the features extracted by the generator, the discriminator needs to be able to efficiently capture the key features of the notes of the music score image and filter the effects of irrelevant noise in the note sequences. In this paper, the discriminator incorporates BiGRU to obtain the original features of the input note sequences.

$$T_d = BiGRU(x_d, t_{d-1}; \vartheta_d) \tag{17}$$

where $\vartheta_d$ represents the model parameters of BiGRU.

Subsequently, the attention mechanism is adopted to compute the attention value of the output potential characteristic values at each time step, which represents the importance

of different features, and can help the discriminator to capture more key features and ignore the influence of irrelevant features.

$$\tilde{Y}_d = T_d \cdot softmax\big(\alpha^T \cdot \tanh(T_d)\big) \tag{18}$$

where $T_d$ represents the original features of the input, $\alpha$ represents the attention parameters, and $\alpha^T$ is the transpose of $\alpha$. To enhance the utilization of different features of the input information, the discriminator serves as an input to the subsequent model by fusing the attention output features with the original output features. Ultimately, the fused features output from the model are linearized using a fully connected layer to output the final binary classification detection result.

For the binary classification problem, a Binary Cross-entropy [28] loss function is used, which is defined as bellow.

$$L_{GAN}(p_i) = -\frac{1}{M}\sum_{i=1}^{M}[y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i)] \tag{19}$$

where $M$ represents the amount of instances, $y_i$ represents the actual note of instance $i$, with a positive type of 1 and a negative type of 0, and $p_i$ represents the probability that instance $i$ is forecasted to be in the positive type, with the following equation.

$$p_d = \frac{\exp(\tilde{Y}_i)}{\sum_{l=1}^{N}\exp(\tilde{Y}_i)} \tag{20}$$

Since this article adds a classification detection module relied on the traditional discriminative network, and adopts the method of training both discriminators and classifiers to make the discriminative network not only have the function of judging truth and falsehood but also have the function of classification, this paper treats the discriminators and classifiers as subparts in the discriminative network, and the discriminators are labeled with the symbol $K$, the classifiers with the symbol $C$, and the discriminative network with the symbol $D$. The objective function of the discriminative network is indicated in Equation (21), which consists of two parts: the loss function $L_K$ for judging truth and falsehood and the loss function $L_C$ for predicting classification.

$$L_D = L_K + L_C \tag{21}$$

$$L_K = L_{GAN}\big(K(x)\big) + L_{GAN}\big(1 - K(G(y,c))\big) \tag{22}$$

$$L_C = L_{GAN}\big(C(x,c)\big) + L_{GAN}\big(C(G(y,c))\big) \tag{23}$$

where $x$ denotes the true sample, which belongs to the distribution $P_{data}(x)$ of the true data, and $K(x)$ denotes the output of inputting $L_K$ into the discriminator. Ideally, the output probability of $L_K$ is 1 for the real sample and 0 for the pseudo-sample, so it is necessary to minimize $L_K \cdot C(G(y,c))$ denotes the probability of the pseudo-sample with respect to the classification label, and $C(x,c)$ denotes the probability of the real sample with respect to the classification label.

## 5. **Performance testing and analysis.**

5.1. **Recognition performance analysis.** For the goal of estimating the validity of the designed method, the dataset used for the experiments is the CAMERA-PRIMUS dataset [29]. This dataset contains 1528 real scores, each score consists of clef, common notes, elevation marks, punctuation marks, rests, etc., and the length of the score is not fixed, and each score contains 3-6 bars. In this paper, 60% of the above dataset is selected for training, 30% for validation and 20% for testing. Comparative experiments are conducted based on the above dataset, for the convenience of analysis, the algorithms in this paper

are denoted as ours, AresNet in literature [17], and AsGAN in literature [20]. All the experiments are carried out in Pytorch programming environment, with the studying rate set to 0.0002, and the Adam optimizer is used for training. The hardware parameters used were 5 GHz Intel Core i7-11390H processor, 16 GB RAM and Windows 10 operating system.

This article adopts a combination of note error rate (SyER), sequence error rate (SeER) [30], accuracy, precision, recall, and F1 performance metrics to conduct comparative experiments on the improved GAN-based music score detection method ours, existing optimized methods AresNet [17], AsGAN [20] in this paper. Table 1 demonstrates the results of the comparison experiments of different methods.

Table 1. Comparative experimental outcome of different methods

| **Method** | **SyER/%** | **SeER/%** | **Precision/%** | **Recall/%** | **F1/%** |
|---|---|---|---|---|---|
| AresNet | 9.72 | 19.4 | 67.1 | 69.5 | 68.3 |
| AsGAN | 4.58 | 13.6 | 78.9 | 81.3 | 80.1 |
| ours | 1.49 | 8.1 | 93.7 | 96.2 | 94.9 |

As can be seen from Table 1, ours is much better than AresNet and AsGAN in all the indexes, AresNet method has big differences in different kinds of notes, and it performs better in the small symbols such as dots, but it can't detect the uncommon symbols and overlapping symbols well. The AsGAN method has better performance in recognizing the types of notes in a simple score, but it is not good enough in detecting the pitch, and the above two methods can not directly convert the music score image into digital format, so the practical application value is low. In contrast, the method of this paper ours binarizes the music score image, adds the attention mechanism in the GAN, captures the detailed characteristics of the music score image, and adds the classification detection module on the basis of the traditional discriminative network, which has a better detection effect. From the results, the SyER and SeER of our method are 1.49% and 8.1%, respectively, which are reduced by 84.67% and 58.25% compared to AresNet method, and 67.47% and 40.44% compared to AsGAN method, respectively. The F1 score of ours in is 94.9%, compared to AresNet method, AsGAN method by 38.95% and 18.48%, respectively. In general, our method exhibits better music score detection performance.

To test the accuracy of various methods in the presence of noise, various levels of Gaussian white noise (signal-to-noise ratios ranging from 40dB to 70dB) were added to the experiment, and their recognition accuracies are indicated in Figure 4. The horizontal coordinate represents the signal-to-noise ratio of the added Gaussian noise, and the vertical coordinate represents the accuracy of the different methods for music score recognition. As can be seen from Figure 4, the recognition accuracy of ours is the highest, reaching more than 90%, and the recognition rate of this method does not decrease significantly when the noise is gradually increased, reflecting that the method is insensitive to the noise. The AresNet method achieves the highest recognition accuracy of 78% under the condition of high signal-to-noise ratio. However, as the noise increases and the signal-to-noise ratio decreases below 50 dB, the accuracy of the AresNet method decreases sharply, and when the signal-to-noise ratio is 40 dB, the recognition accuracy of the AresNet method is only 13%, while the accuracy of the AsGAN method is still stable at about 80%.

5.2. **Analysis of network loss and training elapsed time.** For the goal of further illustrating the effectiveness of the detection accuracy of ours, the changes of the loss

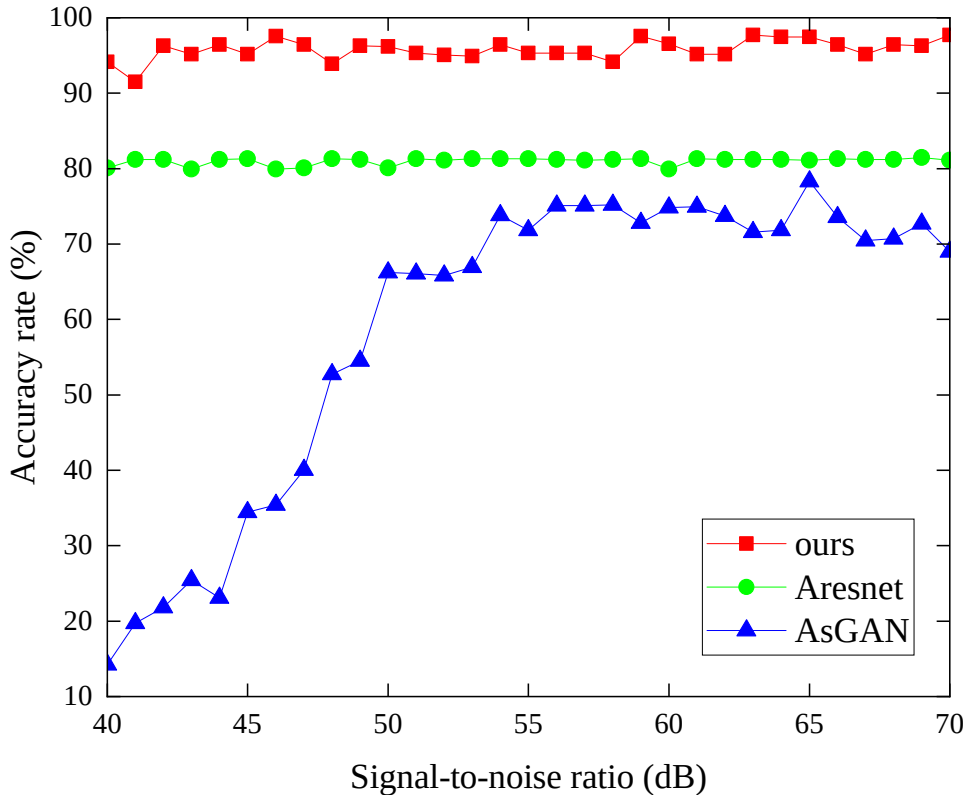


Figure 4. Comparison of the accuracy of various methods in the presence of noise

values with the number of iterations of ours, AresNet and AsGAN methods are compared as shown in Figure 5. It can be seen that the loss value of our method is lower than that of AresNet method and AsGAN method, and keeps a relatively smooth change. Our method's loss value decreases to about 1.5 after 10,000 iterations, whereas the loss value of AsGAN method only decreases to about 2.5 and fluctuates a lot in the later stage, and the loss value of AresNet method decreases to about 2.2 and fluctuates a lot in the middle stage.

To estimate the efficiency of our method in reducing the training time of the models, a training time comparison is performed. Under the same experimental environment, all three models are iterated 64000 times. From the model time curve in Figure 6, it can be seen that the average time consumed by the AsGAN method is about 1.45 minutes per 100 rounds of training, and the total training time is about 15.5 hours, while the average time consumed by the AresNet method is about 1.47 minutes per 100 rounds of training, and the total training time is about 15.7 hours. Our method takes about 0.75 minutes per 100 rounds of training and 8 hours of training. Obviously, the training time of our method is lower than that of the previous two networks, and the training time is about one-half of that of the AsGAN method. This indicates that our method can not only significantly reduce the training time on the basis of guaranteeing the model recognition accuracy, but also improve the model fitting ability and fitting speed.

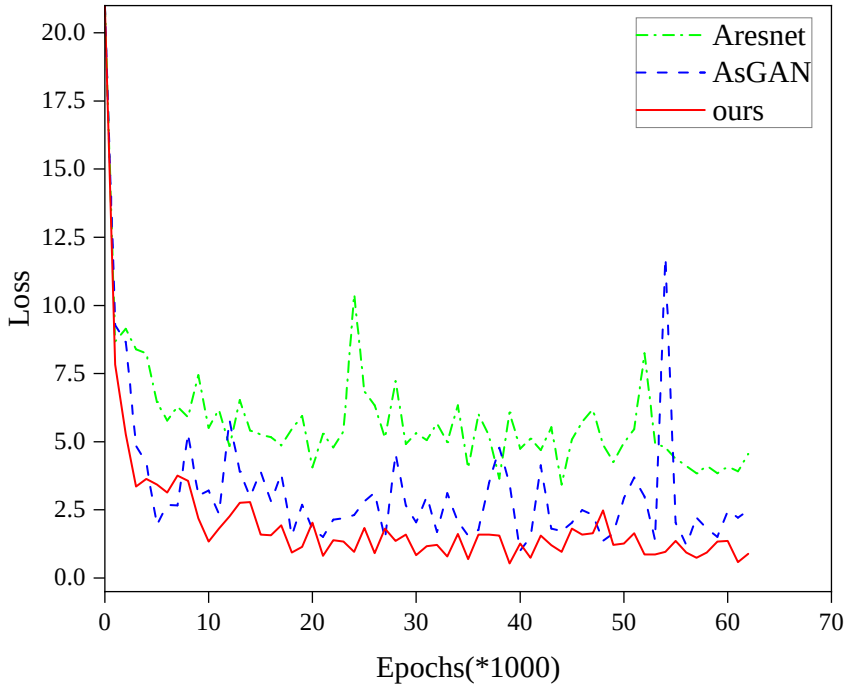


Figure 5. Comparison of loss values for different detection methods.

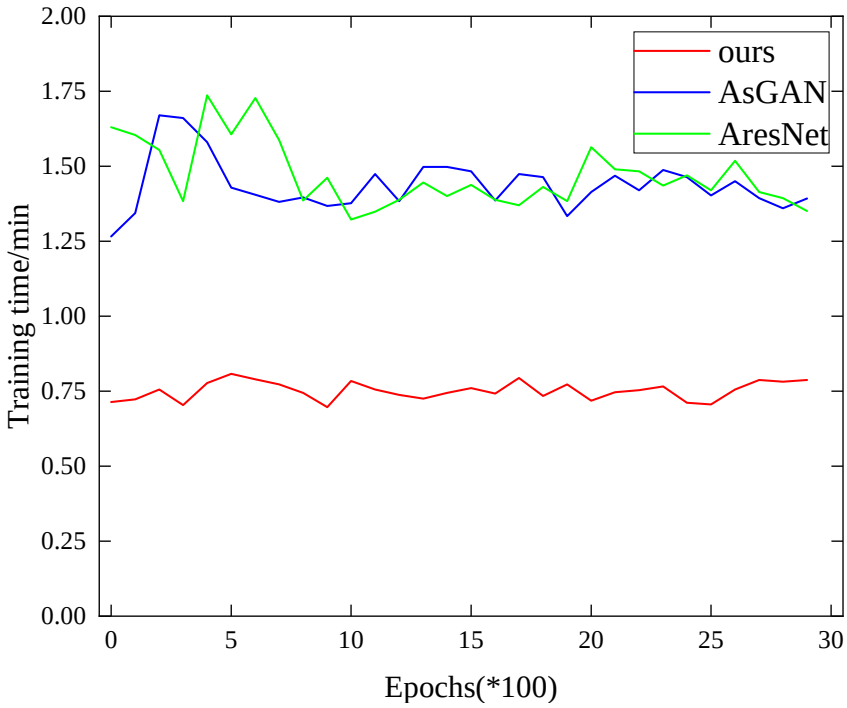


Figure 6. Comparison of training time consumption for different models.

6. **Conclusion.** Intending to the issue of low detection effect of existing music score detection methods, this article suggests a digital music score detection method based on attention mechanism GAN. Firstly, the Wasserstein distance is used to replace the JS dispersion of GAN to optimize the training process of the network, and the fully connected module is introduced to replace the basic module of the generator to improve the feature extraction capability of the network. Secondly, gray scale conversion, sharpening and binarization are performed on the spectral images to remove irrelevant features in

the process of image recognition. To further enhance the detection performance, squeezing and excitation modules are added to the generator of the GAN to explicitly model the interdependence between the feature channels, and then the discriminator fuses the BiGRU to obtain key features, and by fusing the attention output features with the key features as input. Finally, the fused features output from the model are linearized using a fully connected layer to output the final binary classification detection results. The experimental outcome indicates that the suggested method has high detection efficiency compared to the comparison methods.

## REFERENCES


[1] V. Thomas, C. Fremerey, M. Müller, and M. Clausen, "Linking Sheet Music and Audio—Challenges and New Approaches," in *Multimodal Music Processing*, vol. 3, pp. 1–22, 2012.

[2] N. W. Hanrahan, "Hearing the contradictions: Aesthetic experience, music and digitization," *Cultural Sociology*, vol. 12, no. 3, pp. 289–302, 2018.

[3] G. Born and K. Devine, "Music technology, gender, and class: Digitization, educational and social change in Britain," *Twentieth-Century Music*, vol. 12, no. 2, pp. 135–172, 2015.

[4] R. Gaugne, F. Labaune, D. Fontaine, G. L. Cloirec, and V. Gouranton, "From the engraved tablet to the digital tablet, history of a 15th-century music score," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 13, no. 3, pp. 1–18, 2020.

[5] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 173–190, 2012.

[6] M. A. Winget, "Annotations on musical scores by performing musicians: Collaborative models, interactive methods, and music digital library tool development," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 12, pp. 1878–1897, 2008.

[7] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[8] J. Calvo-Zaragoza, J. H. Jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–35, 2020.

[9] Y. Li and W. Zheng, "Emotion recognition and regulation based on stacked sparse auto-encoder network and personalized reconfigurable music," *Mathematics*, vol. 9, no. 6, 593, 2021.

[10] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 765–783, 2020.

[11] J. Calvo-Zaragoza, A. Pertusa, and J. Oncina, "Staff-line detection and removal using a convolutional neural network," *Machine Vision and Applications*, vol. 28, pp. 665–674, 2017.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[13] D. Bisharad and R. H. Laskar, "Music genre recognition using convolutional recurrent neural network architecture," *Expert Systems*, vol. 36, no. 4, e12429, 2019.

[14] X. Han, F. Chen, and J. Ban, "Music emotion recognition based on a neural network with an inception-GRU residual structure," *Electronics*, vol. 12, no. 4, 978, 2023.

[15] Z. Huang, X. Jia, and Y. Guo, "State-of-the-art model for music object recognition with deep learning," *Applied Sciences*, vol. 9, no. 13, 2645, 2019.

[16] C. Cortes, J. A. Ohtola, G. Saggu, and V. P. Ferreira, "Local release of properdin in the cellular microenvironment: role in pattern recognition and amplification of the alternative pathway of complement," *Frontiers in Immunology*, vol. 3, 412, 2013.

[17] S. Rajesh and N. Nalini, "Musical instrument emotion recognition using deep recurrent neural network," *Procedia Computer Science*, vol. 167, pp. 16–25, 2020.

[18] C.-Y. Lee and T.-A. Le, "Identifying faults of rolling element based on persistence spectrum and convolutional neural network with ResNet structure," *IEEE Access*, vol. 9, pp. 78241–78252, 2021.

[19] P. L. Tomaz Neves, J. Fornari, and J. Batista Florindo, "Self-attention generative adversarial networks applied to conditional music generation," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 24419–24430, 2022.

[20] C. Wang, P. Wu, L. Yan, Z. Ye, H. Chen, and H. Ling, "Image classification based on principal component analysis optimized generative adversarial networks," *Multimedia Tools and Applications*, vol. 80, pp. 9687–9701, 2021.
[21] K. Nakamura, T. Nose, Y. Chiba, and A. Ito, "A symbol-level melody completion based on a convolutional neural network with generative adversarial learning," *Journal of Information Processing*, vol. 28, pp. 248–257, 2020.
[22] S. Lattner and J. Nistal, "Stochastic restoration of heavily compressed musical audio using generative adversarial networks," *Electronics*, vol. 10, no. 11, 1349, 2021.
[23] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA: A generative adversarial context encoder for long audio inpainting of music," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 120–131, 2020.
[24] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.
[25] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.
[26] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.
[27] M. Kwaśnicki, "Ten equivalent definitions of the fractional Laplace operator," *Fractional Calculus and Applied Analysis*, vol. 20, no. 1, pp. 7–51, 2017.
[28] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 7, 884, 2019.
[29] M. J. Hall, "Correlation distance and bounds for mutual information," *Entropy*, vol. 15, no. 9, pp. 3698–3713, 2013.
[30] Y. Liu, R. Wu, Y. Wu, L. Luo, and W. Xu, "A Stave-Aware Optical Music Recognition on Monophonic Scores for Camera-Based Scenarios," *Applied Sciences*, vol. 13, no. 16, 9360, 2023.
[31] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation with convolutional recurrent neural networks," *Pattern Recognition Letters*, vol. 128, pp. 115–121, 2019.