ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Piano transcription studies considering music education and the ResNet18 network

Zhe Du[1,*]

[1] Department of Music, Qilu Normal University, Jinan, Shandong, 250200, China

SUMMARY: *With the popularization and development of quality education, more and more people realize the importance of music education to the cultivation of students.In order to promote the intelligent development of music education, this paper conducts research on the piano transcription task.A variety of audio signal processing methods are used for feature extraction, and the ResNet18 network is utilized as an implementation method for piano transcription, with improvements such as incorporating the attention mechanism, improving the activation function and using Dropout.Experiments show that the improved design of the ResNet18 network can play a role in the phenomenon of overfitting, and the Loss values of the training set and validation set of the improved ResNet18 network model are 0.025 and 0.11, respectively, which are smaller than those of the ResNet18 network model.The model in this paper performs optimally in the test set comparison experiments, with higher test metric values than other methods on note start + end point, note start + end point + volume, with 1.91%~5.38% and 1.91%~5.38% improvement, respectively.The proposed improved ResNet18 network model has superior piano transcription performance, which can accurately transcribe the music played by the performer to assist teaching work.*

KEYWORDS: *ResNet18 network; Dropout; Attention mechanism; Piano transcription; Music education*

1 Introduction

Music brings endless joy and tranquility to people, standing the test of time as a unique symbol throughout history while continuously evolving and growing. In the past, musicians used sheet music and instruments to create a variety of artistic treasures that have been passed down to this day, giving rise to musical genres represented by wind and string instruments. With the continuous advancement of technology, digital music has emerged and played a pivotal role in the creation and dissemination of music.

Automatic Music Transcription (AMT) technology is the process of converting raw audio information into musical notation or MIDI data [1]. MIDI represents the detection of pitch, start and end times of notes, and other related information in audio [2]. AMT technology can assist in music creation, making it easier to reproduce music. AMT systems play an important role in many scenarios. In the field of music education, AMT can automatically identify notes played on instruments, provide timely evaluation and feedback, point out deficiencies and errors in learners' performances, thereby improving learning efficiency and reducing the burden on families and teachers [3-6]. In the field of music information retrieval, AMT can translate raw audio signals into intuitive musical notation, facilitating the extraction of deeper

*ddyjy1030@163.com
https://doi.org/10.65102/is20261096 Published: 10 June, 2026

semantic features and improving music retrieval efficiency [7-9]. In music notation annotation, many audio files require manual notation, which is labor-intensive and time-consuming. AMT can quickly detect audio and generate corresponding sheet music, significantly reducing annotation time [10-12]. In music appreciation, it can be used to visualize music content, presenting musical performances in various forms, making them more accessible to the general public [13, 14]. In summary, AMT systems have a wide range of applications and diverse uses, so developing a high-performance, stable AMT system is of significant research importance.

As a typical representative of polyphonic instruments, piano music transcription has been a classic research problem. Literature [15] investigates piano transcription systems from the perspective of training data, introducing new datasets and data augmentation techniques to effectively address overfitting issues in piano transcription training data, thereby improving the accuracy of the transcription model. Literature [16] introduces a method that fully considers musical structure and transcribes musical audio into piano sheet music. Compared to traditional audio transcription methods that generate each bar independently, the proposed method extracts rhythmic features, melodies, and chords from the input audio recording and reflects them in the sheet music. In addition to audio-based transcription methods, vision-based transcription methods have also emerged as a valuable supplement. Literature [17] combines audio and visual features to establish a piano-specific transcription system. It uses a multi-band signal start point detection method based on specific spectral envelope matching filters for automatic annotation of audio signals, while introducing computer vision methods to enhance the annotation of pure audio piano music. Reference [18] proposes a real-time piano music visual transcription system based on CNN-SVM black-and-white key classification. By analyzing the piano keyboard and the movement trajectories of the performer's fingers, it can achieve real-time transcription of piano music with high accuracy even under non-ideal camera angles. Additionally, some scholars have applied convolutional neural network methods to audiovisual separation tasks and music generation tasks, achieving excellent results. Literature [19] addresses the issue of existing emotion transcription systems in capturing musical temporal details by establishing an artificial neural network model that includes a convolutional neural network and a temporal network to address note offset issues during piano transcription, thereby improving transcription performance. Literature [20] constructed a music melody extraction method based on attention modules and time-domain harmonic mapping, and designed a piano transcription algorithm using this method, which can effectively improve the accuracy of music recognition and annotation, thereby improving piano transcription efficiency.

The study introduces the application of piano transcription in music education, and selects three audio signal processing methods, namely CQT, Meier Transform Frequency, and STFT, for piano music feature extraction, and then utilizes ResNet18 network for piano transcription.Aiming at the problem that this neural network has a large number of neurons, large parameter computation, and is prone to overfitting phenomenon in the piano transcription process, the improvement methods of incorporating the attention mechanism, improving the activation function, and using Dropout are proposed to complete the construction of the piano transcription model based on the improved ResNet18 network.The MAESTEO dataset and the synthetic music dataset are used as the experimental objects to compare the changes of the Loss values of this paper's model and the ResNet18 network model in the training set and validation set experiments, to explore the feasibility of the improvements made in this paper.Comparison experiments of multiple transcription methods and ablation experiments are conducted to explore the practical effect of the proposed piano transcription model based on the improved ResNet18 network.

2 Piano transcription in music education

Automatic Music Transcription (AMT) technology, which is the process of converting raw audio information into music notation or Musical Instrument Digital Interface (MIDI) data, is an important research component of Music Information Retrieval (MIR), and can be used to transcribe music so that it can be easily reproduced when improvising or composing at will.Musical notation generally contains pitch, the start time and the end time of the appearance of the pitch, and the main content of AMT is to predict the events of notes that occur during a certain period of time, and it can be used for tasks such as detecting the start and end times of notes, music score annotation, music teaching and instrument identification.

AMT system has an important role in the field of music education.AMT can automatically recognize the notes in the process of playing musical instruments, give timely evaluation and feedback, point out the learners' deficiencies and mistakes in the process of playing, so as to improve the efficiency of learning and reduce the burden of families and teachers.In music information retrieval, AMT can translate raw audio signals into intuitively represented music symbols, which is conducive to enhancing deep-level semantic features and improving music retrieval efficiency.In terms of music score annotation, many audios need to be manually annotated with music scores, which has high labor and time costs. AMT can quickly detect the audio and generate the corresponding music scores, which greatly reduces the time for music score annotation.In terms of music appreciation, it can be used to visualize the music content and show the music performance content in different forms, which can be easily appreciated by students.In summary, AMT systems have a wide range of application scenarios and diverse uses in music education, so an AMT system with superior performance and stable performance is of great research significance.As a typical representative of polyphonic instruments, piano music transcription has been a classical research problem.In this paper, we study the piano transcription method based on ResNet18 network, which is applied in music education.

3 Piano transcription model based on ResNet18

As a basic part of music education, sight singing is a necessary course for music beginners.In the past, the whole process of sight-singing practice required manual participation, and teachers had to provide note-level evaluation and feedback according to the students' sight-singing performance.Such a teaching mode is inefficient and not conducive to students' independent learning, which will inhibit students' learning motivation and limit the development of related industries.In order to promote the efficient development of music education, this paper proposes a piano transcription model based on ResNet18.

3.1 Data processing and feature extraction

The original audio of the piano has a sampling rate of 44.1KHZ and in the experiment it needs to be downsampled to 16KHZ before training.Regarding the training three types of features CQT, Mel Transform Frequency, and STFT are tried.The three features obtained are two-dimensional time-frequency features, in the process of STFT and Meier transform frequency feature extraction, every 2048 points as a frame using the Hamming window for the calculation, the step size is 512, so a total of 31.25 frames in a second, using the three features to calculate the frequency domain feature dimension size are 252, 252, 1025, respectively.The time domain length is the total number of frames of a single audio.The tags are read and computed by reading each line of the text document, extracting the onset_time

(start time) and offset_time (end time) of each line as well as the pitch, which is converted from the time point to the position of a specific frame, and the result is also a two-dimensional vector, where the dimension of the first dimension corresponds to the time information, which is the total number of frames, as is the case for the corresponding audio feature, and the second dimension is thepitch value, which depends on the total number of pitches set for the piano, and was set to 88 (corresponding to 88 pitches) in the experiments.

The obtained feature data and label data are saved into pickle files using python's pickle module according to the audio name, and in order to speed up the training, the practice of secondary data packing is used, where each pickle file is read cyclically, and the feature files and corresponding labels of each track are added to a list list, and then this list object is directly saved as a newThen we directly save this list object as a new pickle file to achieve the effect of packing and compression. Compared with reading the pickle file in a loop, saving it as a list object for one-time reading can really improve the loading speed of the feature data.

CQT, Mel Transform Frequency, STFT three audio signal processing methods are as follows:

3.1.1 Constant Q-transformations

In order to analyze musical scales more accurately, a time-frequency conversion algorithm of constant Q transform (CQT) is introduced.In CQT, the filter banks satisfy the conditions that the center frequencies are distributed exponentially, the filter bandwidths are different, and the ratio of the center frequencies to the bandwidths is a constant Q. The difference between CQT and Fourier Transform is that the cross-axis frequencies of the CQT spectra are not linear, but exponential.And when the filter window length is changed, it is correlated with the change in the spectral line frequency, resulting in an improvement in performance.Since the CQT has a one-to-one correspondence with the distribution of the scale frequencies, calculating the CQT spectrum of the music signal also gives access to the amplitude values of the music signal at each note frequency.Thus the constant Q transform is able to achieve the same effect of the exponential distribution.

Center frequency division: constant Q invariant means that the ratio of center frequency to bandwidth is constant, the ratio is Q. Q is also called quality factor, which can be calculated by equation (1):

$$Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/b} - 1} \tag{1}$$

where the center frequency of the band is $f_k = f_0 2^{k/b}$, and b is denoted by dividing b b bands over an octave.

Assume that N_k is the length of the window that varies with frequency and that the sampling frequency is f_s.Then the relationship satisfies equation (2):

$$N_k = Q \frac{f_s}{f_k} \tag{2}$$

Then in the constant Q transform, the kth component in the nth frame can be derived from Eq. (3):

$$X_n^{cqt}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) W_{N_k}(n) e^{-\frac{2j\pi Qn}{N_k}} \tag{3}$$

$W_{N_k}(n)$ represents the window function for the kth band.

When the CQT value is obtained for each frame according to the formula, the CQT transformed spectrum can be obtained by taking the amplitude spectra in the segmentation window and splicing them together.

3.1.2 Meier transforms

The human ear does not pick up sound in all frequency bands, and is like a filter bank that focuses only on certain specific frequency components, letting certain frequency signals pass through and simply ignoring certain frequency signals that it does not want to perceive.The human ear usually receives sound information in the range of 20 to 20,000 Hz.That is, the human auditory nerve perceives frequencies selectively, without linear correlation.It can be understood in this way that when the human ear hears a sound at 1000 Hz, and when the frequency of the sound suddenly increases by 2000 Hz, the human ear does not perceive that a large fluctuation has occurred, and will only perceive that the frequency has increased partially, and will not realize that the frequency has doubled.The relationship between the human ear's perception of frequency and the Meier frequency is changed to a linear relationship through the Meier scale.In the above example, when the frequency of the sound is doubled, the perception of the human ear upon hearing it is also doubled.The conversion of a common frequency scale corresponding to the Mel scale is shown in the formula:

$$f_{mel} = 2595 * \log_{10}(1 + \frac{f}{700}) \tag{4}$$

where f is the frequency value of the audio at the frequency scale and f_{mel} is the frequency value of the audio at the Meier scale.

3.1.3 Short-time Fourier transforms

The Fourier transform transforms the information corresponding to the time domain into the frequency domain, allowing for subsequent analysis of frequency characteristics in the frequency domain.Although a spectrogram can show the distribution of frequencies more clearly, the element of how long a frequency is maintained is missing.For similar extended problems such as analyzing precisely when a frequency begins and ends, the Fourier transform becomes difficult to cope with.In order to solve such problems, the short-time Fourier transform (STFT) was created to link information in the time and frequency domains.

The STFT first splits the signal into frames, divides the whole signal into sub-signals of the same size, and then performs a windowing operation, i.e. multiplying each sub-signal by a non-zero window function, and then stacks the results according to a dimension, and the stacked spectrograms clearly show the time-frequency information characteristics.

The short-time Fourier transform form can be expressed by Equation (5):

$$X(t, f) = \int_{-\infty}^{\infty} \omega(t-\tau) x(\tau) e^{-j2\Pi f \tau} dt \tag{5}$$

where $\omega(t)$ is the window function and $X(t, \omega)$ is the Fourier transform of $\omega(t-\tau)x(\tau)$ .As t changes, the window function is displaced on the time axis. After

$\omega(t-\tau)x(\tau)$, only the intercepted portion of the window function is left for the final Fourier transform of the signal, and the result is a complex function representing the magnitude and phase of the signal as it changes with time and frequency.

3.2 ResNet18 network model and its improvement

3.2.1 ResNet18 network modeling

Based on the fact that traditional convolutional neural network models are prone to overfitting or model degradation as the number of network layers increases, the residual network ResNet has been proposed, which can increase the number of model layers to the level of a hundred layers.ResNet no longer lets the next layer directly fit to the underlying mapping, and the core idea is that: since traditional CNN models are basically set as shallow to prevent overfitting, when their model accuracy reaches saturation, several constant mapping layers with output equal to input are added.The core idea is: since the traditional CNN model is basically set as a shallow layer in order to prevent overfitting phenomenon, when the accuracy of the model reaches saturation, several constant mapping layers with output equal to input are added, i.e., the information is sent to the deeper layer of the neural network through the "jumping board" by skipping one layer or more layers of connection to increase the depth of the network without increasing the error.Thus, ResNet improves the accuracy of its neural network by adding more layers to the conventional CNN model.

The structure of the basic unit of the ResNet model, shown in Figure 1, is similar to a "shortcut" in an electrical circuit, called a shortcut connection, which means that it can skip one or more layers of connections and send information to deeper layers of the neural network.

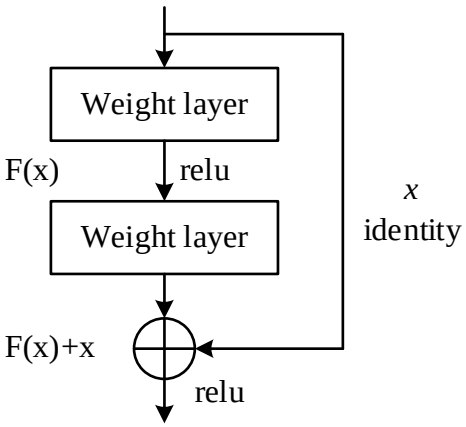


Figure 1: Structure Diagram of ResNet Basic Unit

The shortcut connection is equivalent to simply performing the equivalent mapping without generating additional parameters or increasing computational complexity, and its entire network can still be trained by end-to-end backpropagation.From a mathematical point of view, the shortcut connection process is analyzed as follows:

$$y = F(x, W_i) + x \tag{6}$$

$$F = W_2\sigma(W, x) \tag{7}$$

The addition of $F(x)$ to x is element-by-element, and if the two have different dimensions, a linear mapping needs to be performed on x to match the dimensions:

$$y = F(x, W_i) + W_{s^x} \tag{8}$$

where x is the input, y is the output, $F(x, W_i)$ is the residual function, σ is the ReLU activation function, and w_2 is the weight.

There are five main forms of ResNet: ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152.The network used in this paper is ResNet18, which consists of 17 convolutional layers and one fully connected layer.

In experiments using the ResNet18 network as a piano transcription model, the number of neurons often reaches hundreds, and the number of weights in the same layer of the network can reach the level of 100,000 or more.In machine learning and deep learning if the model has too many parameters and not enough samples, the model is prone to overfitting, and the ResNet18 network is also improved in order to enhance the piano transcription effect of the ResNet18 network.

3.2.2 Integration of attention mechanisms

Convolutional Attention Mechanism (CAM) in neural networks has been widely used in the field of music recognition, and the Convolutional Attention Module CBAM) consists of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM).These two sub-modules pay attention to the channel information and spatial information respectively, and reconstruct a series of features in the middle of the network again, which highlights some important features and suppresses other general features, which can achieve the ultimate purpose of enhancing the audio recognition effect.

The specific process of attentional reconstruction is shown in Equation (4), where F denotes the feature map of a network layer in the network model, $M_e(F)$ denotes the one-dimensional channel attentional reconstruction of F using CAM, and F' denotes the feature map that has undergone channel attentional reconstruction, and $M_s(F)$ denotes the two-dimensional spatial attentional reconstruction of F' usingSAM for 2D spatial attention reconstruction of F', F' denotes the output feature map that combines both channel and spatial aspects of attention, and $\otimes$ denotes element-by-element multiplication.For the 3D feature map $F \in R^{C \times H \times W}$ of a certain network layer in a convolutional neural network CNN, one-dimensional channel attention feature map M_e and two-dimensional spatial attention feature map M_s are successively inferred from F and multiplied element by element, respectively, to finally arrive at the output feature map that is equal to F .dimension output feature map F'':

$$\begin{cases} F' = M_c(F) \otimes F \\ F' = M_s(F') \otimes F' \end{cases} \tag{9}$$

For the residual network ResNet18, the most critical location for extracting features should be in each Bottleneck.Considering all the considerations, in this paper, CBAM is fused between each ©layer, and the main reason for doing so is that the residual network model ResNet18 has already completed the feature extraction in each Bottleneck, and the CBAM will then perform the attentional reconfiguration hereThe main reason for doing so is that the

residual network model ResNet18 has completed feature extraction in each Bottleneck, and CBAM can play the role of carrying on the previous and the next.

3.2.3 Activation function improvement

The hidden layer activation function of the residual network ResNet uses the linear modified unit function (ReLU) by default, and its mathematical expression is shown in Equation (10):

$$ReLU(x)=\begin{cases}0 & x\le 0\\ x & x>0\end{cases} \tag{10}$$

That is, when x is greater than 0 the function value is equal to x itself, and when x is less than or equal to 0 its value is all 0.

The ReLU function is very simple and effective, its advantage is high computational efficiency with fast convergence, but the biggest disadvantage is the neuron necrosis problem.In order to solve the above problems, the paper uses the parameterized linear correction unit function (PReLU) to replace the original ReLU activation function in the residual network ResNet, as shown in Equation (11):

$$PReLU(x)=\begin{cases}x & x>0\\ a_i x & x\le 0\end{cases} \tag{11}$$

Usually the parameter a_i is relatively small and generally lies between 0 and 1.In the negative interval $[0,-\infty)$, PReLU also has a small learnable slope, which not only retains the advantages of the ReLU function in the original positive interval $[0,\infty)$ and strengthens the expressive ability of the model, but also avoids the "DeadReLU" mentioned above." problem mentioned above can be avoided at the same time.

3.2.4 Using Dropout

Dropout does this by randomly and temporarily logically isolating a portion of neurons from the network according to a certain probability during the training of a deep neural network model.For stochastic gradient descent, because of the random and temporary isolation strategy, each small batch is training a different network, but after the completion of the model training, each neuron will not be absent, and will participate in the actual reasoning.Dropout can reduce the dependence of the neurons on each other, and more effectively alleviate the phenomenon of overfitting of network models, and to a certain extent can achieve the goal of reducing structural risk and improving the generalization ability of the model.To a certain extent, it can achieve the effect of reducing structural risk and improving model generalization ability.

In order to further prevent model overfitting and improve the robustness of the network, the paper tries to use the Dropout strategy between avgpool and fc in the output layer of ResNet18, and the relationship between the three after incorporating the Dropout layer is shown in Figure 2.

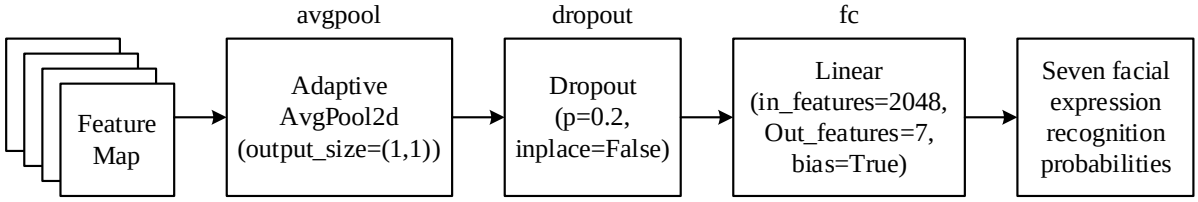


Figure 2: Res Net melding Dropout

4 Experimental results and performance analysis

4.1 Experimental setup

4.1.1 Data entry

The piano music dataset was adopted from MAESTEO___v3.0, which was recorded and organized with the help of a digital piano Disklavier with 1,276 tracks and a total duration of 198.7 hours.The synthesized piano music part uses six sf2 piano sound sources that are publicly available for free on the web, and is synthesized by the fluidsynth method.

4.1.2 Training parameters

The experiments were conducted in half precision using adam strategy training with a batch size of 12 and a learning rate decaying from 0.003, updated every 5 steps, with pre-training and fine-tuning performed 30,000 times each.The CPU part of the training in the presence of synthesized music input runs 12 threads simultaneously to provide synthesized data.

4.1.3 Evaluation methodology

The final transcription results were evaluated using the current FI evaluation method common to piano music transcription, which establishes a set of reference standards for errors in note start and end points, volume, and so on.Since this network adopts the transcription form of note start **-** end points, no frame-level pitch evaluation is performed.

4.2 Experimental results

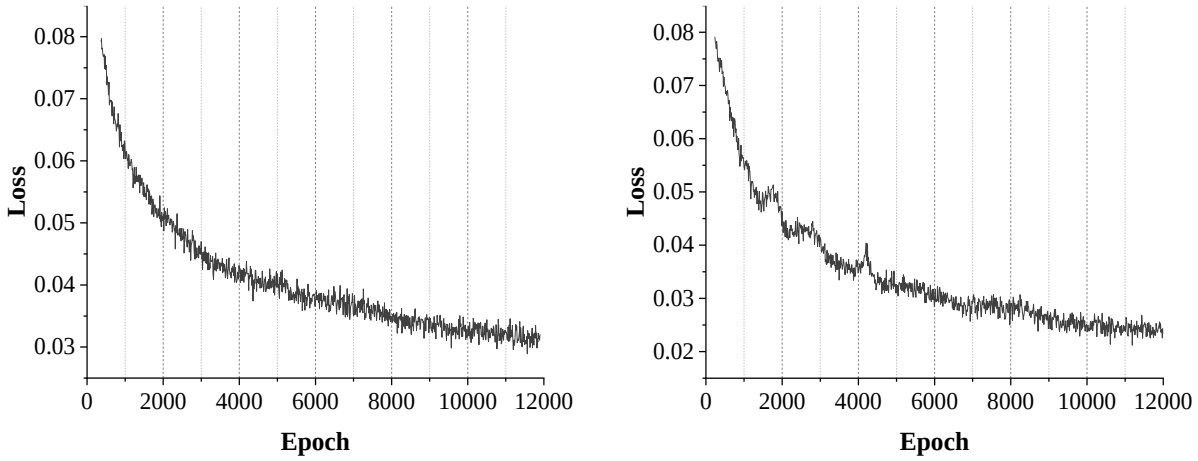
4.2.1 Effect of model improvement

In the task of piano transcription, the effect of the three frequency features is compared under deep neural networks, comparing the differences in the performance of the deep neural networks that undergo the improvement of the fusion attention mechanism, activation function, and Dropout and those that do not undergo the improvement, and multiple sets of comparative experiments are conducted, and the results of the piano transcription experiments are shown in Table 1.The three indexes of the improved ResNet18 network are better than those of the unimproved ResNet18 network, and the accuracy, recall and F-value of the former are above 83%, while the three indexes of the latter are less than 75%.The changes of training set and training set Loss during the training process can be clearly seen by the visualization tool in the experimental process, and the training set Loss trend of the ResNet18 network model and the improved ResNet18 network model are shown in Fig. 3, Fig. (a) and Fig. (b).The validation set Loss trend of the ResNet18 network model and the improved ResNet18 network model are shown in Fig. 4, Fig. (a) and Fig. (b).Loss trend is shown in Figures (a) and (b) of Figure 4.With both Losses essentially no longer changing, the Loss for

the training set of the unimproved deep neural network is around 0.032, while the Loss for the validation set is as high as 0.17. In contrast, the Loss for the training set of the improved ResNet18 network model is around 0.025, while the Loss for the test set is around 0.11.This shows that the difference between the Losses of the test set and the training set of the improved ResNet18 network is much smaller, and that the unimproved ResNet18 network with too many neurons is more computationally intensive and prone to overfitting.

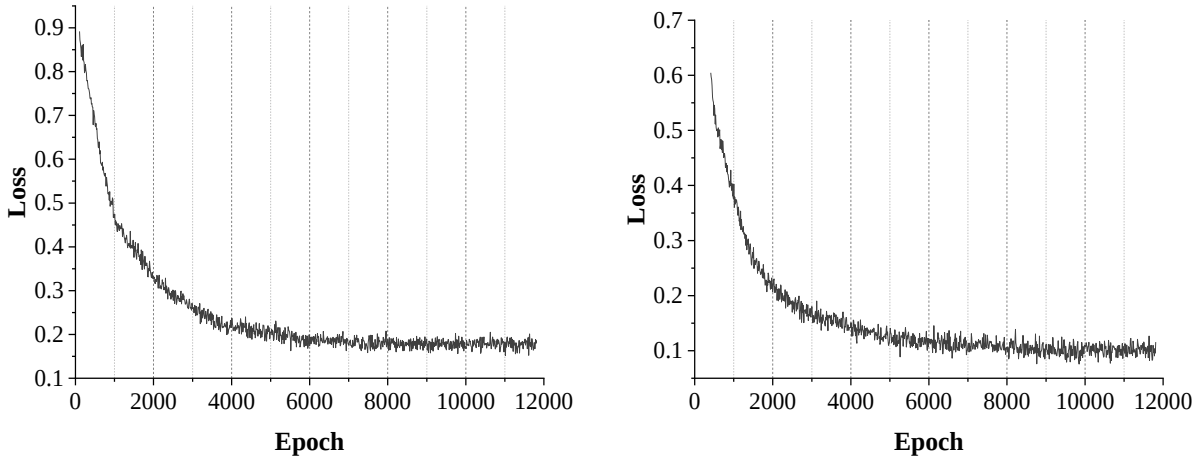
Table 1: Piano transcription experiment results

Model	Precision/%	Recall/%	F-measure/%
ResNet18-CQT	74.41	73.85	74.13
ResNet18-CQT (improved)	83.53	88.56	85.97
ResNet18-Mel (improved)	87.85	91.76	89.76
ResNet18-STFT (improved)	85.31	88.52	86.89



(a)ResNet18 (b)Improved- ResNet18

Figure 3: ResNet18 network model training set loss trend



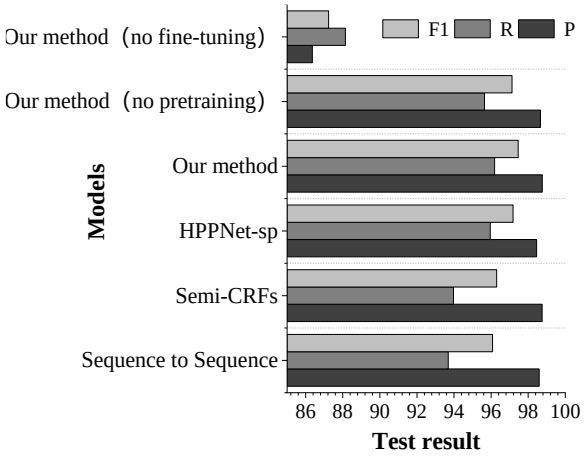
(a)ResNet18 (b)Improved- ResNet18

Figure 4: ResNet18 network model verification set loss trend

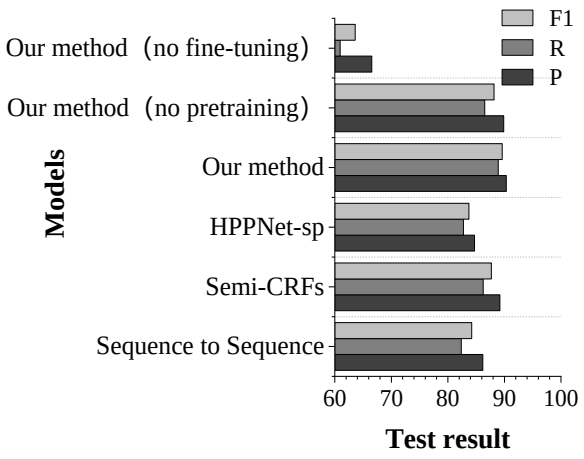
4.2.2 Test set results

Comparison experiments with other piano transcription methods, Fig. 5 shows the

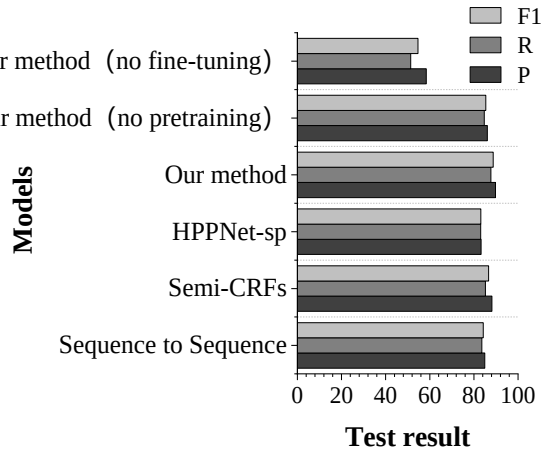
experimental results of this experiment on MAESTEO__v3 test set, Fig. (a) ~ Fig. (c) show the test results of each piano transcription model for note start, note start + end point, note start + end point + volume, respectively.The recognition rate of note starting point of each method is high with little difference, while the recognition with note ending point has significant difference, and the F1 result of this paper's method with ending point reaches 89.61%, which is 1.91%~5.38% higher than other methods.In the test of note start + end point + volume, the F1 value of this paper's method reaches 88.74%, which is improved by 2.08%~5.59% than other methods.Comparison of the final network with the version without pre-training (without synthetic music) shows that even the simplest synthetic music scheme can improve the current piano music transcription performance.The version without fine-tuning demonstrates the magnitude of the difference between this synthesized music and the MAESTEO dataset, with recognition particularly evident at the end point.With the current piano music dataset still small in size and relevant reviews in real music composition yet to be refined, there is still room to explore the application of synthesized music.



(a)Note starting point



(b)Note starting point+ End point



(c)Note starting point+ End point+ Volume

Figure 5: MAESTEO test result

4.2.3 Ablation experiments

(1) Input features

Table 2 lists the results of music transcription with different input features, where the training set is trained using synthesized music and MAESTEO___v3 is used as the validation

set.With a sufficiently large model and sufficient amount of data, the more sufficient information contained in the input features, the better the training results are. The F1 values of piano transcription with the three features input simultaneously are 82.15% and 92.11% for the frame-level pitch and note starting point of the training set, and 87.45% and 88.80% for the validation set, which are better than the other feature input methods.Also the more severe the overfitting of the network relative to the MAESTEO dataset.Currently, the Meier transform map is commonly used in the field of music transcription as an input feature, which is concise and effective.However, in this experiment the amount of data provided by the synthesized music is very sufficient, allowing the network to grasp more details, so the feature input form of CQT, Mel Transform plus STFT is used to make the relevant conclusions more universal.

Table 2: effect of input features on results

Input feature	Training set						
	Training loss	Height of frame			Note starting point		
		P	R	F1	P	R	F1
CQT	0.341	82.27	78.28	80.23	96.31	87.57	91.73
CQT +Mel	0.339	82.65	77.21	79.84	96.08	87.61	91.65
CQT +Mel+ STFT	**0.333**	83.61	80.74	**82.15**	97.36	87.39	**92.11**
Input feature	Verification set						
	Verification loss	Height of frame			Note starting point		
		P	R	F1	P	R	F1
CQT	**0.338**	62.03	87.15	72.48	91.46	85.32	88.28
CQT +Mel	0.406	57.61	86.19	69.06	91.77	85.49	88.52
CQT +Mel+ STFT	0.409	87.33	87.58	**87.45**	90.68	86.99	**88.80**

(2) Input duration

Figure 6 demonstrates the relationship between the input music clip duration on the note starting point and the frame-level pitch recognition results for its associated experiments with a Mel-transformed spectrum for the inputs, and a relatively simplified structure for the overall network to allow for quick experiments.From the results, the recognition results are relatively best when the duration of the input segment is around 10s, which is related to the overall duration distribution of the notes, and at this time, the frame-level pitch F1 value and the note starting point F1 value are 88.86% and 96.44%, respectively.

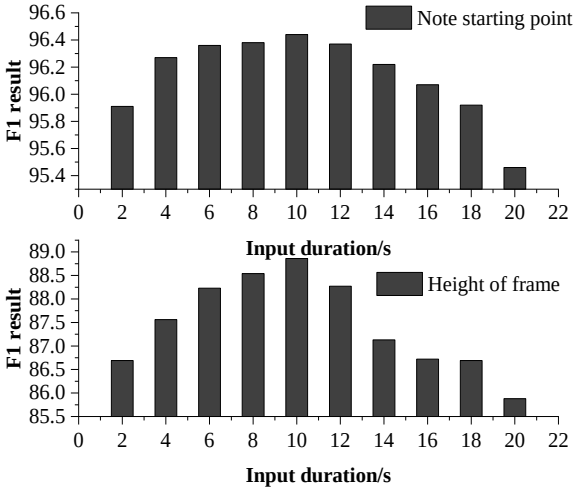


Figure 6: The impact of the input length on the result (MAESTEO verification set)

5 Conclusion

In this paper, we use ResNet18 network to explore the piano transcription, combining multiple audio feature extraction methods, and propose an improved piano transcription model for ResNet18 network.The MAESTEO dataset and the synthesized music dataset are used as samples for experiments, and the main research results are as follows:

(1) The Loss values of the improved ResNet18 network model in this paper are 0.025 and 0.11 in the training and validation sets, respectively, while the Loss values of the ResNet18 network model are 0.032 and 0.17, respectively.The Loss values of this paper's model are much smaller in the two datasets, and the overfitting phenomenon can be improved, which suggests that incorporating attention in the ResNet18 networkmechanism, improved activation function and the effectiveness of the improved design using Dropout.

(2) The test results of this paper's method are not much different from those of other methods on the note starting point, and the F1 results at the note starting point + ending point are improved by 1.91% to 5.38%, and the F1 results at the note starting point + ending point + volume are improved by 1.91% to 5.38%, which demonstrates the superiority of the proposed improved ResNet18 network model for piano transcription.

(3) In the ablation experiments, the more input features, the better the model is for piano transcription, and when all three features are input, the model's F1 values for the frame-level pitch and the note starting point are 82.15%~92.11%.In addition, when the audio input duration is 10s, the model's transcription is the best, and the model's F1 values for frame-level pitch and note starting point reach the maximum, which are 88.86% and 96.44%, respectively.

An effective a piano transcription method was realized in this study, which can be used to assist music education and help music education to be intelligent and efficient.In the future, it is necessary to explore methods that can make the model's effect further improved, such as in the laboratory's previous research, it was proposed to comprehensively determine the transcription results by jointly judging the pitch and the note start and stop time in order to improve the effect of automatic music transcription.

References

[1] Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2018). Automatic music transcription: An overview. IEEE Signal Processing Magazine, 36(1), 20-30.

[2] Grossman, A., & Grossman, J. (2020). Automatic music transcription: generating midi from audio. Reports of CS230 Deep Learning, Lecture at Stanford University, 1-6.

[3] Vaca, K., Gajjar, A., & Yang, X. (2019, July). Real-time automatic music transcription (AMT) with Zync FPGA. In 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI) (pp. 378-384). IEEE.

[4] Wu, Y. T., Chen, B., & Su, L. (2020). Multi-instrument automatic music transcription with self-attention-based instance segmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2796-2809.

[5] Bhattarai, B., & Lee, J. (2023). A comprehensive review on music transcription. Applied Sciences, 13(21), 11882.

[6] McLeod, A., Schramm, R., Steedman, M., & Benetos, E. (2017). Automatic

transcription of polyphonic vocal music. Applied Sciences, 7(12), 1285.

[7] Román, M. A., Pertusa, A., & Calvo-Zaragoza, J. (2020). Data representations for audio-to-score monophonic music transcription. Expert Systems with Applications, 162, 113769.

[8] Ycart, A., Liu, L., Benetos, E., & Pearce, M. (2020). Investigating the perceptual validity of evaluation metrics for automatic piano music transcription. Transactions of the international society for music information retrieval.

[9] Rizzi, A., Antonelli, M., & Luzi, M. (2017). Instrument learning and sparse NMD for automatic polyphonic music transcription. IEEE Transactions on Multimedia, 19(7), 1405-1415.

[10] Kong, Q., Li, B., Song, X., Wan, Y., & Wang, Y. (2021). High-resolution piano transcription with pedals by regressing onset and offset times. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3707-3717.

[11] Yu, Y. (2024, September). Research on Automatic Piano Music Transcription Algorithm Based on CNN. In Proceedings of the 2024 2nd International Conference on Internet of Things and Cloud Computing Technology (pp. 199-204).

[12] Liang, Y., & Pan, F. (2023). Study of Automatic Piano Transcription Algorithms based on the Polyphonic Properties of Piano Audio. IEIE Transactions on Smart Processing & Computing, 12(5), 412-418.

[13] Li, Y., Wang, X., Wu, R., Xu, W., & Cheng, W. (2024). A Two-Stage Audio-Visual Fusion Piano Transcription Model Based on the Attention Mechanism. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 3618-3630.

[14] Shibata, K., Nakamura, E., & Yoshii, K. (2021). Non-local musical statistics as guides for audio-to-score piano transcription. Information Sciences, 566, 262-280.

[15] Edwards, D., Dixon, S., Benetos, E., Maezawa, A., & Kusaka, Y. (2024). A data-driven analysis of robust automatic piano transcription. IEEE Signal Processing Letters, 31, 681-685.

[16] Takamori, H., Nakatsuka, T., Fukayama, S., Goto, M., & Morishima, S. (2019). Audio-based automatic generation of a piano reduction score by considering the musical structure. In MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25 (pp. 169-181). Springer International Publishing.

[17] Wan, Y., Wang, X., Zhou, R., & Yan, Y. (2015). Automatic piano music transcription using audio-visual features. Chinese Journal of Electronics, 24(3), 596-603.

[18] Akbari, M., Liang, J., & Cheng, H. (2018). A real-time system for online learning-based visual transcription of piano music. Multimedia tools and applications, 77, 25513-25535.

[19] Chen, Z., Zhou, Y., & Wu, H. (2025, April). Research on an automatic transcription

method for piano music based on artificial neural networks. In Fourth International Conference on Computer Technology, Information Engineering, and Electron Materials (CTIEEM 2024) (Vol. 13561, pp. 205-213). SPIE.

[20] Wu, Q., & Yu, T. (2025). Piano Transcription Using Temporal Harmonic Diagram and Transfer Window Attention in Self-Attention Networks. Informatica, 49(5), 133-146.