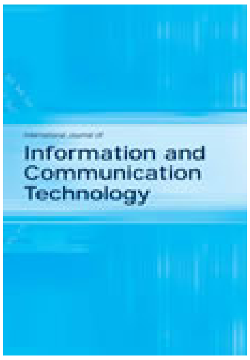
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Multi-instrument polyphonic automatic transcription method combining gated recurrent units and DeepLabv3+ model

Xiaochen Ye

DOI: 10.1504/IJICT.2026.10075866

Article History:

Received:	17 October 2025
Last revised:	26 November 2025
Accepted:	01 December 2025
Published online:	04 February 2026

Multi-instrument polyphonic automatic transcription method combining gated recurrent units and DeepLabv3+ model

Xiaochen Ye

Music College,
Neijiang Normal University,
Neijiang, 641100, China
Email: xiaochen_ye@outlook.com

Abstract: A multi-instrument polyphonic automatic transcription method integrating bidirectional gated recurrent units and an improved Deeplabv3+ network is proposed to enhance transcription accuracy under complex audio conditions. A pre-separation module first performs source separation and denoising. Frequency-harmonic composite features are then extracted, and temporal dependencies are modelled using a gated recurrent network, followed by lightweight decoding for note onset localisation and instrument classification. Experiments show that the proposed model achieves 92.8%, 91.5%, and 92.1% accuracy, recall, and F1 on the training set, and 91.2%, 88.7%, and 89.9% on the test set, surpassing baseline methods. In mixed-instrument scenarios, the model attains an average F1 of 83.65% and 88.3% note recognition accuracy, improving piano-violin transcription by 7%. The method offers high precision and robustness for polyphonic transcription, providing a practical foundation for intelligent music analysis and automatic orchestration.

Keywords: multi-instrument polyphonic auto-transcription; DeepLabv3+ network; bi-directional gated loop unit; audio feature extraction; preamplifier separation.

Reference to this paper should be made as follows: Ye, X. (2026) ‘Multi-instrument polyphonic automatic transcription method combining gated recurrent units and DeepLabv3+ model’, *Int. J. Information and Communication Technology*, Vol. 27, No. 4, pp.69–90.

Biographical notes: Xiaochen Ye holds a Doctoral in String Performance from the Saint Petersburg State Conservatory. He is currently a Lecturer at the School of Music, Neijiang Normal University, teaching courses including history of western music, basic music theory, harmony, chamber music and string instrument performance. He has published a number of academic papers in various domestic journals. He studied under several Russian Meritorious Artists and served as their teaching assistant. He was invited to conduct academic visits and participate in master classes in multiple countries. With extensive orchestral performance experience, she has worked with many prestigious orchestras such as the Symphony Orchestra of the Mariinsky Theatre and the Saint Petersburg Philharmonic Orchestra.

1 Background

With the continuous development of artificial intelligence technology in the field of music information processing, automatic music transcription (ADT) has gradually become one of the key tasks in intelligent music analysis and digital content generation (Wang et al., 2024). In the context of multi-instrument polyphony, problems such as overlapping note frequencies, mutual interference of different instrument timbres and background noise make it much more difficult to model the transcription system (Lee and Jeong, 2023). A number of scholars at home and abroad have conducted research on ADT systems. For example, Cahyaningtyas et al. (2023) proposed an ADT method based on deep learning. The method first performed segmentation by parameter optimisation, and subsequently the extracted spectral features were fed into a classification model. Experimental results indicated that the method achieved 77.42%, 86.97% and 82.87% multi-objective optimisation scores on different datasets.

Park et al. (2023) proposed a singing melody transcription model based on a sequence-to-sequence transformer. The model represented the melody as a monophonic sequence, used overlapping decoding to ensure context continuity, and enhanced the generalisation ability of the model through pitch enhancement and noisy data cleaning. The results of ablation experiments indicated that the model outperformed existing schemes in all evaluation metrics. Velazquez Lopez et al. (2022) proposed a piano music transcription system based on an improved non-negative matrix decomposition, which enhanced the Fourier spectrogram visual representation by a novel cochlear filter. System evaluation showed that the scheme achieved higher accuracy in the task of transcribing polyphonic piano music, validating the effectiveness of auditory feature filtering. Lee and Lee (2024) used fast Fourier transform (FFT) and short-time Fourier transform (STFT) to extract musical bass notes. This study identifies frequency separated fundamental frequencies by Hamming window FFT. In a simple polyphonic music test, the word error rate was 3.13% and the character error rate was 6.25%, verifying the effectiveness of the method.

Transcription task has also made significant development, and convolutional neural networks are widely used in the field of image semantic segmentation (ISS) because of their ability to effectively capture spatio-temporal features in audio (Preethi and Mamatha, 2023). The DeepLabv3+ network efficiently extracts multi-scale features through the collaboration of the atrous spatial pyramid pooling (ASPP) module and the decoder. It performs well in the field of ISS (Chen et al., 2024). Ji et al. (2022) proposed a semantic segmentation method based on multilayer feature fusion. The method improved the accuracy of semantic segmentation by introducing a flexible and lightweight extrusion excitation module into the spatial pyramid pool (SPP) network. The enhanced multilayer feature fusion structure may greatly increase the accuracy of semantic fusion, according to experimental results. Yang et al. (2022) proposed an ISS method based on deep neural networks, which extracted pixel-level and image-level features through convolutional structures, and fused the features after refining them using upsampling. Experimental results revealed that the method outperformed the comparison method in terms of performance and operation speed.

However, the standard structure of convolutional networks suffers from a large number of parameters and insufficiently targeted feature learning in audio transcription tasks, which makes it difficult to be directly applied to multi-instrument scenarios. In contrast, bidirectional gated recurrent unit (BiGRU) can efficiently model the timing dependence of sequence data. It also has the advantages of low computational overhead and stable gradient transfer (Wang et al., 2023). Suresh Kumar and Rajan (2023) proposed a transformer-based multimodal music mood classification system and compared the performance with BiGRU-based system. Additionally, they analysed the performance of other advanced methods. The outcomes showed that the transformer model achieved higher accuracy than the single-layer attention-based multi-modal system using BiGRU, with a maximum accuracy of 77.94%. Mohamed and Yassine (2023) employed multimodal feature learning using a Siamese network to learn distance measures between audiovisual features. The study used the GRU-Attention network to learn sequential semantic and spatial location information, and then combined principal component analysis with the Python program Tsfresh to extract features from the power spectral density of audio streams. Experimental results indicated that incorporating audio features significantly improved the F1-score and gradient detection performance.

In summary, although existing research has made some progress in note recognition, timbre differentiation, and temporal modelling, the traditional segmentation classification-based AMT method in Cahyaningtyas et al. (2023) is insufficient for modelling the temporal evolution process. The transformer-based schemes in Park et al. (2023) and Suresh Kumar and Rajan (2023) have certain advantages in global sequence modelling, but their ability to recover local time-frequency structures is limited, and attention dilution is prone to occur under long sequence input conditions. Lee and Lee (2024) relies on the combination of FFT and STFT features for statistical estimation methods, which are suitable for low complexity polyphonic scenes, but are susceptible to note overlap interference in complex multi-track environments. In contrast, BiGRU can bidirectionally model the temporal evolution of musical notes, enhance rhythm and structural boundary recognition, and has advantages in smaller parameter scale and stable gradient transfer, making it more suitable for real-time scenes. The improved DeepLabv3+ achieves fine-grained reconstruction of spectral semantics through multi-scale dilated convolution and staged deconvolution, effectively compensating for the shortcomings of traditional convolutional networks in long-term frequency coupling modelling.

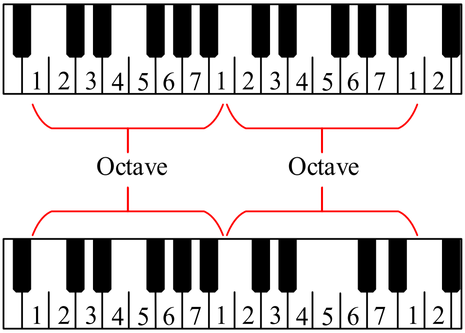
In light of this, the study proposes a multi-instrument polyphonic automatic transcription method that combines BiGRU with an improved DeepLabv3+ network. It innovatively introduces a pre-source separation structure to preprocess mixed audio, further reducing the interference of background noise on modelling accuracy. The study placed BiGRU before spectral feature integration, forming an encoding order of ‘temporal first, semantic second’ to involve temporal prior information in subsequent spatial structure optimisation, and completing feature domain alignment through 1 × 1 convolutional mapping instead of simple parallel connections or hard stacking. The combination of the two forms a collaborative mechanism of ‘time modelling + spatial refinement’, which is more suitable for dealing with complex tone spectral line crossing and pronunciation structure coupling problems compared to single convolutional neural networks, transformers and other structures, demonstrating higher theoretical adaptability and modelling hierarchy advantages. The study aims to integrate temporal modelling and spatial decoding capabilities in order to achieve the high-precision transcription and classification of multi-instrument signals. This will be accomplished through noise reduction preprocessing, composite spectral feature construction, and multi-level deconvolution structures.

2 Improving the DeepLabv3+ network architecture and multi-instrument audio transcription method

2.1 Music input and feature construction

The goal of ADT is to convert audio signals into corresponding digital symbols, with the key being to identify the pitch, start time, and end time of notes (Edwards et al., 2023). The piano, for example, has 88 keys covering a frequency range of 27.5 Hz to 4,186 Hz, and produces sound through the vibration of strings when played. Each note can be viewed as the superposition of multiple sine waves, forming a harmonic structure. A note consists of a transient and a steady state, with the onset typically determined by identifying energy changes. For example, the piano produces a sound that rises rapidly, enters a stable phase, and then decays gradually. This sound has a rich tone and long reverberation. Modern music uses the 12-tone equal temperament (12-TET) system, which divides an octave into 12 semitones, standardising the pitch system. The distribution of same-named notes across different octaves on the keyboard provides a reference for automatic recognition, as shown in Figure 1 (Wang, 2023).

Figure 1 The distribution of the same named notes on the keyboard (see online version for colours)



In Figure 1, the 12-TET divides the octave into 12 semitones, with the smallest pitch interval being the adjacent semitone. According to the 12-TET, the frequency relationship of each note in the piano range is shown in equation (1) (Yi et al., 2024).

$$f_n = f_0 \cdot 2^{\frac{n}{12}} \tag{1}$$

In equation (1), f_n is the frequency of the n^{th} note. f_0 represents the fundamental frequency of the leftmost note on the piano keyboard, typically set to 27.5 Hz. $\frac{n}{12}$ denotes the number of semitones above the reference pitch, with each octave consisting of 12 semitones. In time-frequency analysis, the STFT has insufficient resolution in the high-frequency range due to its fixed window length. Additionally, its linear spectrum does not match the exponential characteristics of the piano signal, resulting in poor analysis performance (Simonetta et al., 2022). In contrast, the constant-Q transform (CQT) uses a logarithmic frequency axis and a dynamically changing window length.

This design accurately reflects the audio characteristics of the piano and reduces the number of model parameters due to its lower spectral dimension. These features make the CQT more suitable for piano audio analysis. CQT is defined as shown in equation (2) (Wang and Dai, 2025).

$$X(k)=\sum_{n=0}^{N_k-1} x(n)\cdot w_k(n)\cdot e^{-j2\pi\frac{\eta f_k^n}{N_k}} \tag{2}$$

In equation (2), $X(k)$ represents the CQT coefficient corresponding to the k frequency component. η is the sampling rate. f_k represents the k frequency point. $x(n)$ represents the original discrete audio signal. $w_k(n)$ represents the window function corresponding to frequency f_k. N_k represents the window length corresponding to frequency f_k. $e^{-j2\pi\frac{f_s f_k^n}{N_k}}$ represents a complex sine wave with frequency f_k. f_k is defined as shown in equation (3).

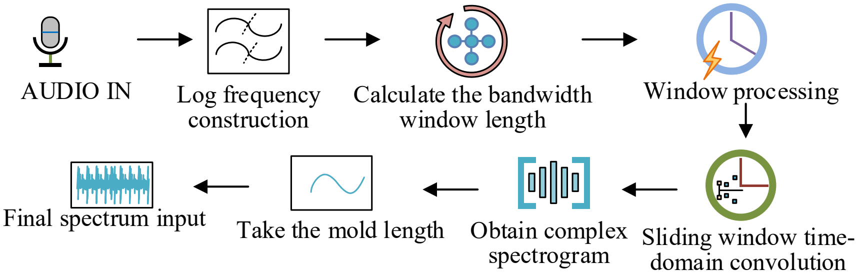
$$f_k = f_{\min}\cdot 2^{\frac{K}{B}} \tag{3}$$

In equation (3), $f_{\min}$ represents the lowest frequency. B represents the number of frequencies within each octave. K represents the total number of frequency bands. The time-frequency spectrum expression is displayed in equation (4) (Spoorthy and Koolagudi, 2024).

$$X(t,k)\in R^{T\times K} \tag{4}$$

In equation (4), $X(t, k)$ is the two-dimensional CQT spectrum. T is the quantity of time frames. $X(t, k)$ can be input into the DeepLabv3+ network for automatic transcription, pitch recognition, note start and end detection, and other tasks. The CQT process is shown in Figure 2 (Peng, 2023).

Figure 2 CQT process (see online version for colours)

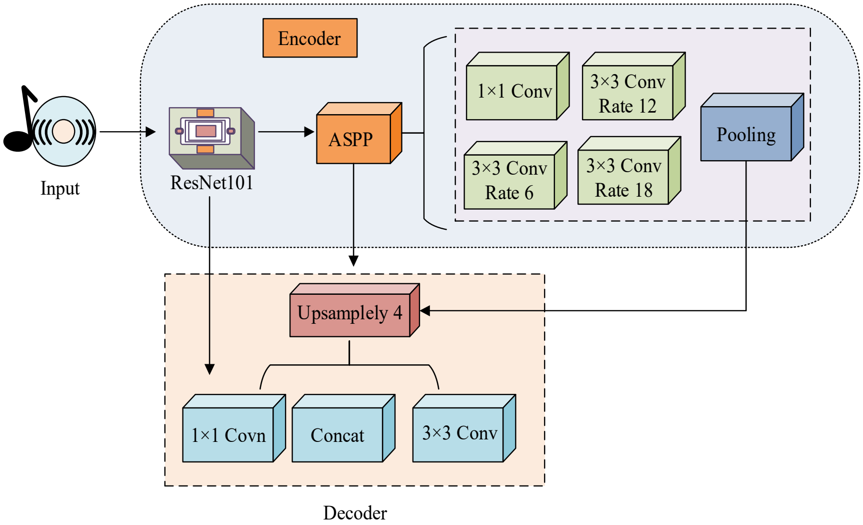

AUDIO IN
Log frequency construction
Calculate the bandwidth window length
Window processing
Sliding window time-domain convolution
Obtain complex spectrogram
Take the mold length
Final spectrum input

In Figure 2, the CQT calculation process mainly includes signal preprocessing, frequency axis construction, windowing, and frequency response calculation. First, the original audio signal is standardised, and then the frequency axis is constructed on a logarithmic scale based on the set minimum frequency and the number of frequencies per octave. Next, for each frequency component f_k, the window length N_k is calculated based on its corresponding period length, and a corresponding window function is designed for it. Next, a sliding window is applied to the signal to extract local segments. These segments are weighted by the window function and multiplied by the complex exponential basis function N_k at the corresponding frequency. This process extracts the response intensity of that frequency component. By repeating this process for all frequency components and all time frames, the resulting $X(t, k)$ can be used for subsequent note recognition and model input.

2.2 *Multi-instrument polyphonic transcription model based on GRU-DeepLabv3+*

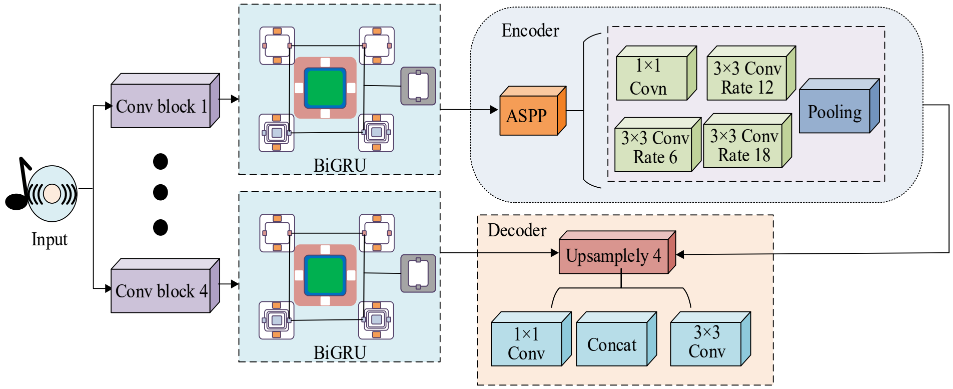
After constructing and preprocessing the spectral features of the audio signal, it is necessary to further design an efficient deep learning network structure to achieve accurate note recognition and transcription. The long-term dependencies in audio signals are better suited for modelling in the time-frequency domain, and music processing can be viewed as a semantic segmentation task on time-frequency images. The Deeplab series of networks capture multi-scale features through dilated convolutions, with DeepLabv3+ demonstrating the best segmentation performance. Its network architecture is shown in Figure 3 (Wang and Dong, 2024).

Figure 3 DeepLabv3+ network structure (see online version for colours)



In Figure 3, the model adopts an encoder-decoder structure. In the encoder part, ResNet101 deep residual network is first used as the feature extraction backbone to perform multi-level feature representation on the input spectrogram (Wu et al., 2022). Then, the ASPP module is introduced to achieve multi-scale feature fusion by using hollow convolution operations with different sampling rates in parallel. Finally, the dimension is reduced by 1 × 1 convolution. The decoder performs deconvolution processing on the encoded features and fuses them with the backbone network features to ultimately output the classification results. Since audio signal transcription requires modelling temporal context information, the DeepLabv3+ network has limited temporal modelling capabilities and an excessive number of model parameters, which can easily lead to feature information loss and other issues. To address this, the study proposes a BiGRU-DeepLabv3+ network structure. The structure is shown in Figure 4.

Figure 4 BiGRU-DeepLabv3+ network structure (see online version for colours)



In Figure 4, the model uses the extracted feature map (FM) as input. Before inputting into the BiGRU module, take the spectral vector corresponding to each time frame as an input sequence element. After modelling the sequence along the timeline using BiGRU, the dimension of the output temporal feature matrix becomes T × 256, where 256 is the output dimension of the bidirectional structure. First, it performs preliminary temporal modelling on the input feature sequence. Through BiGRU, it filters and remembers the information generated by each layer, effectively reducing redundant information interference and improving the temporal correlation of feature expression. To ensure that the output of BiGRU is consistent with the dimensions of the subsequent DeepLabv3+ encoder structure, the model adds a 1 × 1 convolutional linear mapping layer after BiGRU, converts the T × 256 temporal features into a form that can be fused with the convolutional backbone features, and forms a T × 256 × 1 FM through spatial expansion to input into the subsequent encoding module.

When entering the encoder stage, it is first concatenated directly with the spatial features output by the DeepLabv3+ encoder along the channel dimension to form a joint feature tensor. Subsequently, the model adopts a streamlined main feature extraction structure. It replaces the ResNet101 backbone network of the original DeepLabv3+ with a submodule containing only four convolutions. Each module adopts a '3 × 3 convolution + 1 × 1 shortcut connection' structure, with input channels set to 64, 128, 256, and 512 in sequence. This reduces the parameter size and speeds up training. Following the main trunk extraction module, the model incorporates an enhanced ASPP module. This improves the ability to capture local and mesoscale semantic features by setting a smaller hole rate. This avoids the sparsity of features caused by a large hole rate. The hole rate of the ASPP module is adjusted from (6, 12, 18) to (3, 6, 9) to adapt to the high-frequency local density features of the music spectrogram structure, while adding a 1 × 1 convolution branch for global semantic compensation. Additionally, to improve the spatial resolution of high-level semantic information, the network introduces a deconvolution structure during the decoding stage, effectively achieving FM upsampling and boundary restoration. The final output stage performs feature fusion via 1 × 1 convolution to generate prediction results. The BiGRU module structure is shown in Figure 5 (Xu et al., 2023).

In Figure 5, the BiGRU module adopts a two-layer bidirectional structure, with 128 hidden units in each direction, resulting in a bidirectional output dimension of 256. x_t

and h_t are the input vector and hidden state (HS) at the current time step. $h_{t-1}$is the hidden state of the previous time step. y_t is the final HS of the output. Z_t is the update gate. R_t is the reset gate. H_t represents the candidate HS. σ is the Sigmoid activation function. BiGRU enhances semantic information capture capabilities by modelling historical and future dependencies in parallel through two GRU submodules: forward and backward. Although ResNet101 alleviates the vanishing gradient problem, its deep structure contains redundancy, with many network layers serving only to prevent model degradation. Therefore, the study streamlines the backbone network of DeepLabv3+ to four convolutional modules and adopts a layer design from shallow to deep to enhance feature learning capabilities. The specific structure is shown in Figure 6 (Todjro and Mensah, 2023).

Figure 5 BiGRU module structure (see online version for colours)

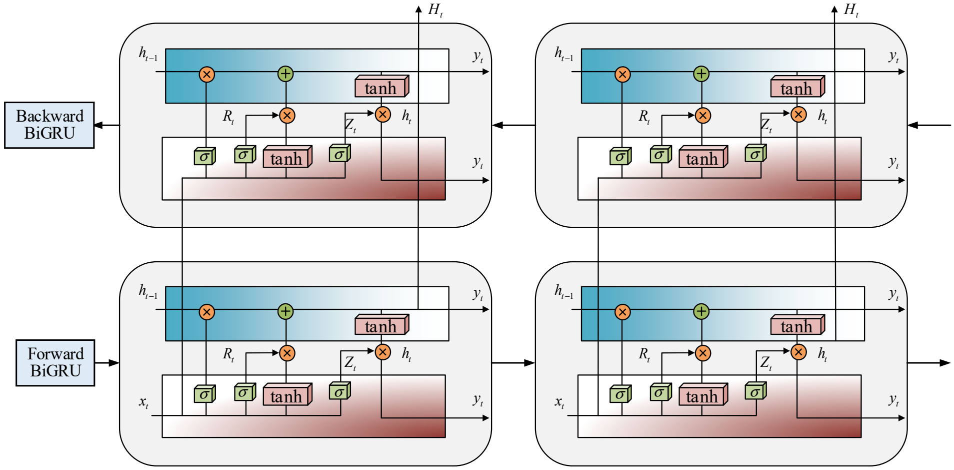
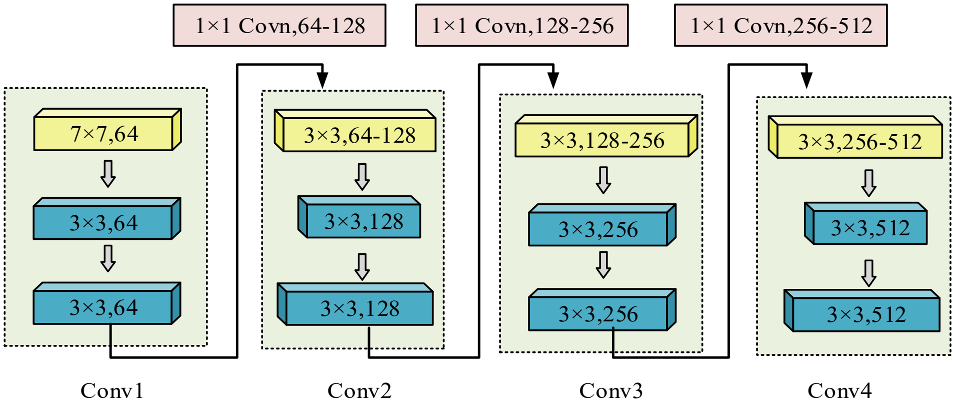


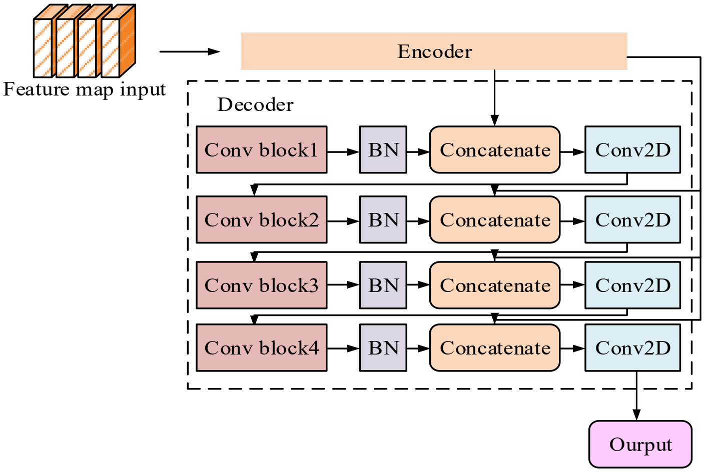
Figure 6 Internal structure of convolutional block (see online version for colours)



In Figure 6, the Conv1 module first uses a 7 × 7 convolution to expand the number of input channels from 3 to 64, followed by a pooling operation to reduce the spatial dimension. The Conv2 module consists of multiple stacked 3 × 3 convolutions, with channel dimensions ranging from 64 to 128, and uses a 1 × 1 convolution to align the dimensions, facilitating residual connections. The Conv3 and Conv4 modules have similar structures, expanding the number of channels from 128 to 256, and then from 256 to 512, respectively. Each stage contains multiple 3 × 3 convolution layers and uses 1 × 1 convolution to construct constant channel or dimension-increasing shortcut paths to form a residual structure. Based on the optimisation of the encoder structure, to further improve the accuracy of polyphonic information recovery for multiple instruments, the study improves the original DeepLabv3+ decoder. The original decoder uses a two-stage 4× upsampling deconvolution module, which can easily cause the loss of high-frequency detail information.

Compared to directly using high magnification upsampling, this progressive approach can achieve gradual reconstruction of local semantic features at each stage, which helps preserve high-frequency details and improve the smoothness of boundary restoration. At the same time, the intermediate FM size of the four-stage decoding structure matches more closely with the time series features generated by BiGRU, avoiding the problem of inconsistent feature scales that may occur in the three-stage design, thereby achieving efficient fusion of spatiotemporal features and improving the accuracy of note boundary localisation. Therefore, in order to balance the semantic information recovery ability and model parameter quantity in the decoding stage, the two-stage upsampling mechanism of the original DeepLabv3+ was optimised to a four level deconvolution structure after analysing the fuzzy characteristics of musical score boundaries and the time sensitivity of note start and end positions in multi-scale audio spectra. The structure is shown in Figure 7 (Xu and He, 2023).

Figure 7 Internal structure of decoder (see online version for colours)



In Figure 7, the improved decoder uses a 2× upsampling deconvolution module to gradually restore the FM size. After each module, a batch normalisation (BN) layer is added, combined with convolutional BN ReLU lightweight mapping to achieve feature compression and spatial alignment, replacing the standard 4× interpolation method and reducing boundary blurring caused by high magnification upsampling (Peng et al., 2023). By combining the compressed features from the encoder with the reconstructed features from the decoder and applying two-dimensional convolution processing, the feature interaction capability is effectively enhanced. The polyphonic transcription output of the BiGRU-DeepLabv3+ network includes instrument categories, note pitches, and timing information. For noisy multi-instrument audio, a preprocessing module based on convolutional time-frequency spectrum separation network is first used for noise reduction processing. This module adopts the MultiConv TPsnet sound source separation scheme, and its core structure consists of depthwise separable convolution and time-frequency domain attention mechanism. By performing multi-scale convolution scanning on the STFT converted spectrogram, it achieves adaptive separation of the main voice and background noise. Separating the network to output a filtering matrix in the form of a spectral mask, multiplying it with the original audio to restore the target sound source spectrum, can improve the signal-to-noise ratio (SNR) in the harmonic region. Afterwards, transcription is carried out through the BiGRU-DeepLabv3+ network. First, this process extracts combined frequency and periodicity (CFP) features and analyses the fundamental frequency and harmonic relationships to obtain the harmonic structure. Then, it performs instrument identification and note transcription based on the unique timbre characteristics of different instruments (Talwar et al., 2023). CFP features combine the advantages of STFT and periodic spectra, obtaining spectra through STFT as shown in equation (5) (Luo et al., 2022).

$$X_{STFT}(t, f) = \sum_{n=0}^{N-1} x(n) \cdot w(n-t) \cdot e^{-j2\pi \frac{f_n}{N}} \tag{5}$$

In equation (5), $X_{STFT}(t, f)$ represents the spectral coefficient at time frame t and frequency f. $w(n - f)$ represents the window function used to extract the local signal segment centred at t. Equation (6) provides the calculation of the periodic spectrum.

$$C(t, q) = \left| IFFT\left(\log\left|X(t, f)\right|\right)\right|^2 \tag{6}$$

In equation (6), q represents the sampling point on the cycle axis. $C(t, q)$ represents the intensity of the periodic component in the signal. CFP feature fusion is shown in equation (7).

$$F_{CFP}(t, f) = X(t, f) \cdot C(t, \tau_f) \tag{7}$$

In equation (7), $F_{CFP}(t, f)$ is the CFP fusion feature. τ_f represents the period position corresponding to the frequency, which is used to remap the period information to the frequency domain. The total loss function (LF) is a multi-task weighted combination form, as shown in equation (8).

$$L = \lambda_1 L_{onset} + \lambda_2 L_{offset} + \lambda_3 L_{pitch} + \lambda_4 L_{inst} \tag{8}$$

In equation (8), L is the total LF. L_{onset} is the note onset detection loss. L_{offset} represents the note offset detection loss. L_{pitch} represents the pitch classification loss. L_{inst} represents the instrument recognition loss. λ_1, λ_2, λ_3, and λ_4 are the corresponding weight coefficients for each sub-loss.

3 Results

3.1 BiGRU-DeepLabv3+ performance testing

To validate the overall performance of the proposed BiGRU-DeepLabv3+ model in multi-instrument polyphonic transcription tasks, the study is based on an Ubuntu 20.04 system, NVIDIA RTX 3090 GPU, Intel i9-12900K processor, and 32 GB of memory. All model modules are implemented using the Python 3.10 and PyTorch 2.0 deep learning frameworks, and the Slakh2100 multi-instrument synthesis dataset is selected as the primary test dataset. Slakh2100 covers a variety of instruments, including piano, guitar, strings, and percussion, and supports note-level annotation and instrument labeling, making it suitable for comprehensive evaluation of polyphonic transcription tasks. First, a systematic ablation experiment is designed to validate the contribution of each module to the overall model performance. This experiment primarily analyses the effectiveness of the BiGRU temporal modelling structure, the ASPP module in DeepLabv3+, the deconvolution structure in the decoder, and the pre-source separation module. Onset detection F1-score (Onset-F1), Offset Detection F1-score (Offset-F1) and pitch recognition F1-score (Pitch-F1) are used. Table 1 displays the findings.

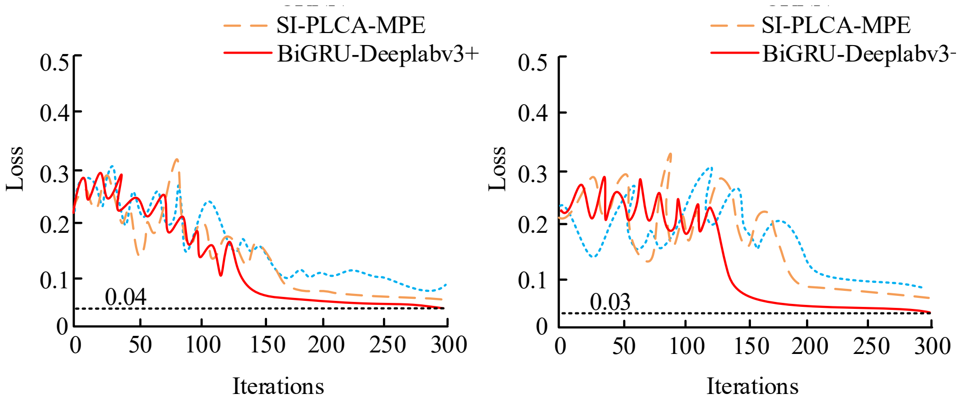
Table 1 Results of ablation experiment

Model	*Onset-F1 (%)*	*Offset-F1 (%)*	*Pitch-F1 (%)*
BiGRU-DeepLabv3+	89.15	86.45	82.71
Remove BiGRU module	84.25	81.07	77.66
Remove ASPP structure	85.62	82.36	78.14
Remove the decoder deconvolution structure	86.13	83.58	79.48
Remove MultiConv-Tpsnet preprocessing	83.78	80.54	76.36

In Table 1, the complete BiGRU-DeepLabv3+ model achieves the highest performance across all three metrics. Specifically, the note onset detection accuracy reaches 89.15%, the note offset detection accuracy is 86.45%, and the pitch recognition accuracy is 82.71%, demonstrating excellent overall recognition capabilities. When the GRU module is removed, the note onset detection accuracy and note offset detection accuracy decrease to 84.25% and 81.07%. This indicates that BiGRU plays a crucial role in modelling note time series contexts, particularly in note boundary detection. After removing the ASPP structure, the pitch recognition accuracy decreases from 82.71% to 78.14%, indicating that ASPP plays a key role in extracting multi-scale frequency features, which helps improve pitch recognition accuracy. Removing the deconvolution structure from the decoder results in a relatively small decline in various indicators, but still shows a certain degree of performance degradation. This indicates that deconvolution positively affects spatial feature restoration and boundary refinement. In addition, when the pre-processing step of separating the pre-source module is removed, the overall performance of the model declines most significantly, with the three metrics decreasing by approximately 5 percentage points each. This validates the important supporting role of the pre-processing stage in complex multi-instrument mixed scenarios for noise reduction and source separation in subsequent transcription tasks.

To further evaluate the performance of the proposed BiGRU-DeepLabv3+, the study divides the Slakh2100 dataset into training and test sets at a ratio of 8:2. It selects a transcription algorithm based on convolutional recurrent neural network (CRNN) (Guo and Zhu, 2025), shift-invariant probabilistic latent component analysis multi-pitch estimation (SI-PLCA-MPE) method (Li et al., 2023), and BiGRU-DeepLabv3+ for comparison. Among them, CRNN represents the widely used convolution and temporal joint modelling strategy in the field of audio transcription, which can reflect the basic performance of convolution feature extraction and cyclic temporal modelling in multi-track scenes. The SI-PLCA-MPE method relies on probability graph models to analyse pitch structures, which is representative in dealing with multi-source frequency overlap and stability modelling. Therefore, it can verify the advantages of the proposed method in deep spatiotemporal structure modelling from the perspective of classical statistical modelling. To ensure experimental fairness, the CRNN model adopts a three-layer convolution and single-layer GRU structure, and adjusts the number of hidden units to 128 to maintain consistency with the main network in terms of temporal modelling scale. The SI-PLCA-MPE method is implemented based on the parameter configuration in the original paper, and the number of iterations for spectral decomposition is set to 100 rounds. All models run on a unified hardware platform and the same training set partitioning strategy. When comparing, comprehensive evaluation is conducted using LF, recognition accuracy, precision, recall, and F1-score indicators. The training loss curves for the three methods are shown in Figure 8.

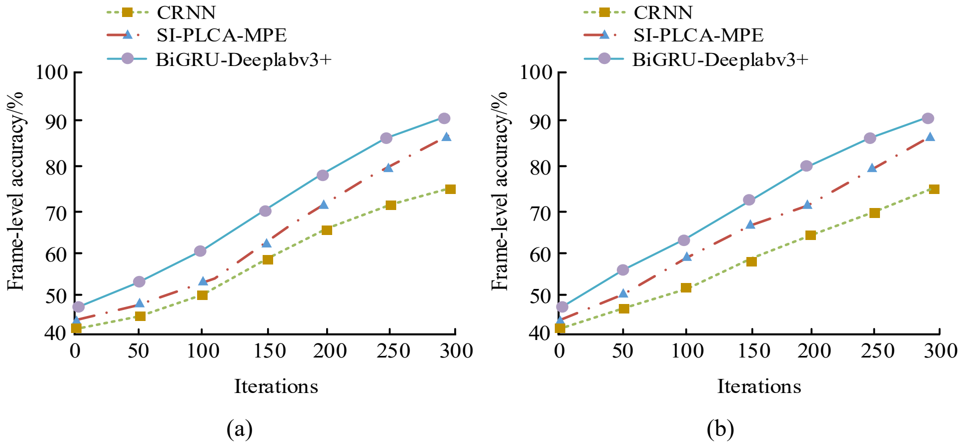
Figure 8 Training loss curve, (a) training set (b) test set (see online version for colours)



(a) (b)

Figures 8(a) and 8(b) show the loss values of the three models on the training set and test set as the quantity of iterations changes. Among them, BiGRU-DeepLabv3+ exhibits better convergence speed and stability. In Figure 8(a), the initial loss of BiGRU-DeepLabv3+ decreases rapidly, stabilises after the 150th iteration, and ultimately reaches the minimum loss value of 0.04 at the 300th iteration, which is significantly lower than CRNN and SI-PLCA-MPE. This indicates that the proposed model has stronger fitting capabilities, smaller training process fluctuations, and a smoother convergence process. In Figure 8(b), the final loss of BiGRU-DeepLabv3+ is 0.03, which is significantly lower than that of CRNN and SI-PLCA-MPE. In conclusion, BiGRU-DeepLabv3+ exhibits a reduced final error during training and a faster rate of convergence. In multi-instrument polyphonic transcription tasks, it also exhibits great stability and robustness during testing, demonstrating its capacity for learning and generalisation. Figure 9 displays the results of several approaches' time frame level recognition accuracy.

Figure 9 Time frame level recognition accuracy results, (a) training set (b) test set (see online version for colours)



The trends in the three models' time frame level recognition accuracy on the training and test sets, respectively, as the number of iterations rises are displayed in Figure 9(a) and Figure 9(b). The BiGRU-DeepLabv3+ model consistently leads during training, as shown in Figure 9(a), and by the 300th iteration, its accuracy has increased to 92%. In contrast, the SI-PLCA-MPE model achieves a final accuracy of 88%, while the CRNN model reaches only 72%. This suggests that the BiGRU-DeepLabv3+ model fits the training set better than other models, especially when it comes to simulating the structural features of note activation frames. Figure 9(b) shows that BiGRU-DeepLabv3+ achieved a frame-level accuracy of 93% on the test set, which is also higher than SI-PLCA-MPE and CRNN. Notably, BiGRU-DeepLabv3+ shows a smoother improvement in accuracy throughout the testing process and consistently outperforms the other two methods. This suggests that while working with unseen samples, the model has superior stability and generalisation skills. The combined data from the two figures further demonstrates the efficacy of BiGRU-DeepLabv3+ in multi-instrument polyphonic transcription tasks by demonstrating both a significant frame-level recognition advantage in the testing stage and a superior learning efficiency in the training stage. Metrics including as accuracy, recall, and F1-score are used in the evaluation. Table 2 displays the findings.

Table 2 Classification performance evaluation

Dataset	*Model*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
Training set	CRNN	82.7	81.9	82.3
	SI-PLCA-MPE	86.4	85.1	85.7
	GRU-DeepLabv3+	92.8	91.5	92.1
Test set	CRNN	80.6	79.2	79.9
	SI-PLCA-MPE	84.5	81.4	82.9
	GRU-DeepLabv3+	91.2	88.7	89.9

In Table 2, BiGRU-DeepLabv3+ significantly outperforms the comparison models in terms of precision, recall, and F1-score on both the training and test sets, demonstrating stronger note recognition capabilities and generalisation performance. When compared against SI-PLCA-MPE and CRNN, BiGRU-DeepLabv3+ maintains its lead on the test set with an F1-score, accuracy rate, and recall rate of 89.9%, 91.2%, and 88.7%. When processing unseen data, the model's recognition performance is more consistent, and its false positive and false negative rates are lower. BiGRU-DeepLabv3+ outperforms SI-PLCA-MPE and CRNN on the training set, achieving an F1-score, precision, and recall rates of 92.1%, 92.8%, and 91.5%, respectively. This illustrates its better modelling benefits and fitting capacity for intricate spectral structures. In conclusion, BiGRU-DeepLabv3+ outperforms probabilistic modelling techniques and conventional convolutional recurrent networks in terms of model stability and note and border recognition accuracy. This makes it a superior option for applications involving multi-instrument polyphonic automated transcription and complicated music signal processing.

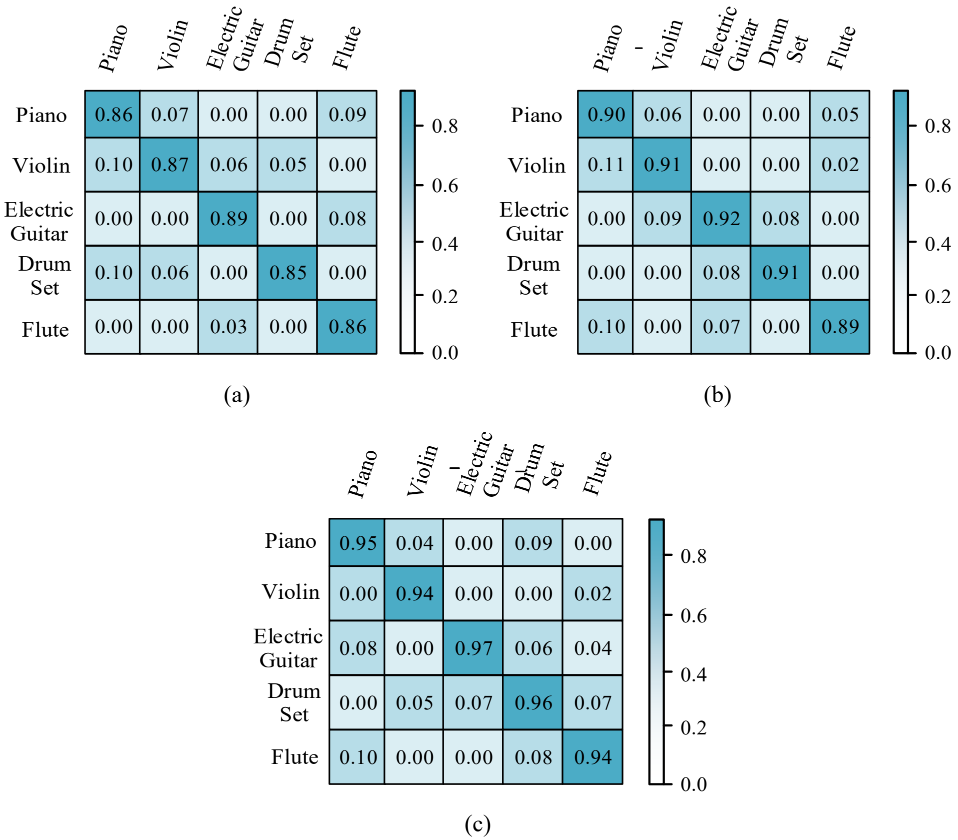
3.2 Analysis of multi-instrument transcription effects

The study performs transcription simulation experiments on polyphonic segments of several instruments to confirm the versatility and applicability of BiGRU-DeepLabv3+ in real complicated musical contexts. The experiment constructs four sets of combined audio data covering typical instruments such as piano, violin, electric guitar, percussion instruments, and flute. Based on the Slakh2100 dataset, mixed audio samples with reverberation, overlap, and style differences are extracted to simulate a real music environment. The recognition accuracy results for multi-instrument mixed segments are shown in Figure 10.

In Figures 10 (a), 10(b), and 10(c) show the confusion matrix results of the three methods in the multi-instrument classification task. In Figure 10(a), the overall recognition accuracy of the CRNN method is relatively low, especially in distinguishing between electric guitars, percussion instruments, and other instruments. The classification accuracy rates for piano and violin are 0.86 and 0.87, respectively. The flute has the highest recognition accuracy value among instruments other than the piano and violin at 0.86, but there is still some degree of confusion. Figure 10(b) shows that the SI-PLCA-MPE method achieves high recognition accuracy for melodic instruments, such as the piano, violin, and electric guitar. Accuracy rates are 0.90, 0.91, and 0.92, respectively, demonstrating the method's advantages in pitch probability modelling. The accuracy rates for percussion instruments and flutes are 0.91 and 0.89, respectively. However, their confusion rates are slightly higher, particularly with minor cross-misclassification between the violin and the electric guitar. This is manifested as mutual interference probabilities of 0.11 and 0.09, respectively.

In Figure 10(c), the overall accuracy of the BiGRU-DeepLabv3+ model is significantly better than the other two methods. The recognition accuracy for piano, violin, and flute is 0.95, 0.94, and 0.94, respectively. The accuracy for electric guitar and percussion instruments reaches 0.97 and 0.96, respectively, which is much higher than CRNN and SI-PLCA-MPE. This indicates that BiGRU-DeepLabv3+ not only demonstrates powerful modelling capabilities in the recognition of melodic instruments, but also exhibits high robustness in the analysis of rhythmic instruments. The study divides five instruments into distinct test groups in order to better assess the model's generalisation abilities in complex instrument combinations. The metrics used to compare the transcription performance of the three approaches in each group are the signal-to-distortion ratio (SDR) and the SNR. Among them, SDR is used to measure the accuracy of the model in reconstructing the target instrument signal during the sound source separation process. The higher the value, the more sufficient the suppression of non target sound sources and noise components, and the transcription results are closer to the real signal. SNR is used to reflect the ratio between the effective instrument signal and background noise in the reconstructed audio, with a high value indicating that the model has stronger anti-interference ability in complex reverberation environments. Table 3 displays the findings.

Figure 10 Identification accuracy results, (a) CRNN (b) SI-PLCA-MPE (c) BiGRU-DeepLabv3+ (see online version for colours)



	Piano	Violin	Electric Guitar	Drum Set	Flute
Piano	0.86	0.07	0.00	0.00	0.09
Violin	0.10	0.87	0.06	0.05	0.00
Electric Guitar	0.00	0.00	0.89	0.00	0.08
Drum Set	0.10	0.06	0.00	0.85	0.00
Flute	0.00	0.00	0.03	0.00	0.86

(a)

	Piano	Violin	Electric Guitar	Drum Set	Flute
Piano	0.90	0.06	0.00	0.00	0.05
Violin	0.11	0.91	0.00	0.00	0.02
Electric Guitar	0.00	0.09	0.92	0.08	0.00
Drum Set	0.00	0.00	0.08	0.91	0.00
Flute	0.10	0.00	0.07	0.00	0.89

(b)

	Piano	Violin	Electric Guitar	Drum Set	Flute
Piano	0.95	0.04	0.00	0.09	0.00
Violin	0.00	0.94	0.00	0.00	0.02
Electric Guitar	0.08	0.00	0.97	0.06	0.04
Drum Set	0.00	0.05	0.07	0.96	0.07
Flute	0.10	0.00	0.00	0.08	0.94

(c)

In Table 3, BiGRU-DeepLabv3+ model outperforms the CRNN and SI-PLCA-MPE methods in both SDR and SNR metrics, indicating its stronger spectral line separation and fidelity characteristics under complex multi-source audio conditions; At the same time, the improvement of F1 score validates the advantages of the model in note structure recovery and time boundary detection, demonstrating high recognition stability and generalisation ability. In the first piano and violin combination, the SDR and SNR of BiGRU-DeepLabv3+ are 12.8 dB and 13.7 dB, respectively, and the F1-score also reaches 88.3%, which is significantly better than that of CRNN and SI-PLCA-MPE. This suggests that the method possesses a good capability of feature recognition in the co-transcription of harmonic instruments.

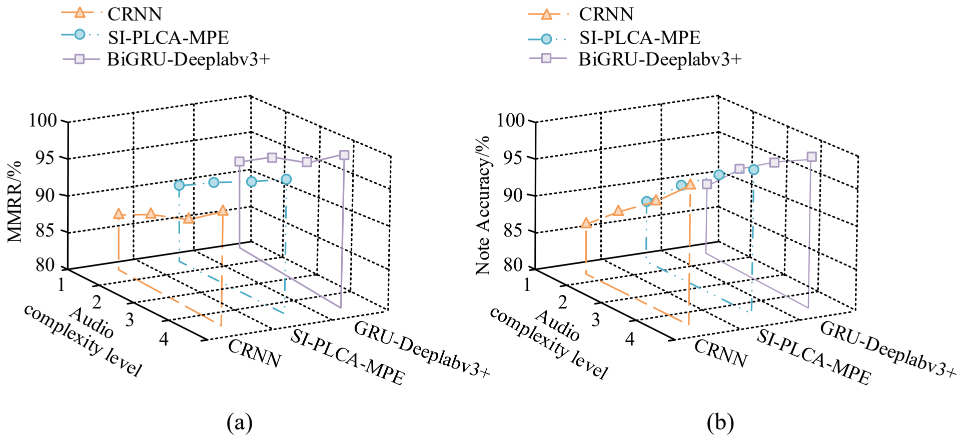
Table 3 Comparison of transcriptional performance among different test groups

Test group number	*Instrument combination*	*Model*	*SDR (dB)*	*SNR (dB)*	*F1-score (%)*
1	Piano + Violin	CRNN	9.6	9.2	83.5
		SI-PLCA-MPE	8.1	10.4	80.7
		GRU-DeepLabv3+	12.8	13.7	88.3
2	Electric guitar + percussion instrument	CRNN	8.7	8.2	77.3
		SI-PLCA-MPE	7.4	9.6	74.2
		GRU-DeepLabv3+	11.9	12.8	84.5
3	Violin and flute	CRNN	8.2	7.9	71.8
		SI-PLCA-MPE	7.1	8.6	69.1
		GRU-DeepLabv3+	10.4	11.3	79.2
4	Piano + electric guitar + percussion instrument	CRNN	9.1	9.1	75.1
		SI-PLCA-MPE	8.3	10.2	71.4
		GRU-DeepLabv3+	12.1	13.3	82.6

In the second group of electric guitar and percussion instruments, the traditional model is prone to miss or confuse the short-time energy peaks due to the strong rhythmic and transient characteristics of this type of instruments. BiGRU-DeepLabv3+, however, is able to maintain a high accuracy rate with an F1-score of 84.5%. The third group is the combination of violin and flute, and the two instruments have close frequency bands, soft timbre and high structural overlap. Among them, the F1-score of BiGRU-DeepLabv3+ is 79.2%, which is significantly better than that of CRNN and SI-PLCA-MPE. This demonstrates the model's good modelling ability for fine-grained spectral features. The fourth group combines three types of instruments with large differences in timbre, piano, electric guitar and percussion, to form a complex polyphonic background. The F1-score of BiGRU-DeepLabv3+ in this scenario is 82.6%, which is an improvement of 7.5 percentage points compared to CRNN. This demonstrates that the model has equally good robustness in the highly diverse mixed-instrument condition.

The study classifies complexity based on the number of instruments that produce sound simultaneously, the density of notes, and the degree of spectral line overlap. Complexity level 1 indicates dominance of a single instrument, with large intervals between notes and no significant overlap; Level 2 represents the simultaneous production of melodic or rhythmic combinations by two instruments, with mild frequency aliasing; Level 3 involves the simultaneous production of three or more instruments, accompanied by continuous notes or rapid rhythm switching, resulting in moderate spectral line crossing; Level 4 corresponds to multi-instrument and multi-segment stacking with obvious harmonic structures, short rhythm intervals, or strong reverberation conditions, making it the most complex scene. The musical instrument digital interface (MIDI), matching reconstruction rate (MMRR) and note accuracy rate are used as indicators. The higher the MMRR, the closer the MIDI structure reconstructed by the model is to the real track, and the more complete the note structure, instrumental hierarchy, and rhythmic reproduction will be; whereas note accuracy is used to measure the overall correctness of the system in terms of note detection. The results are shown in Figure 11.

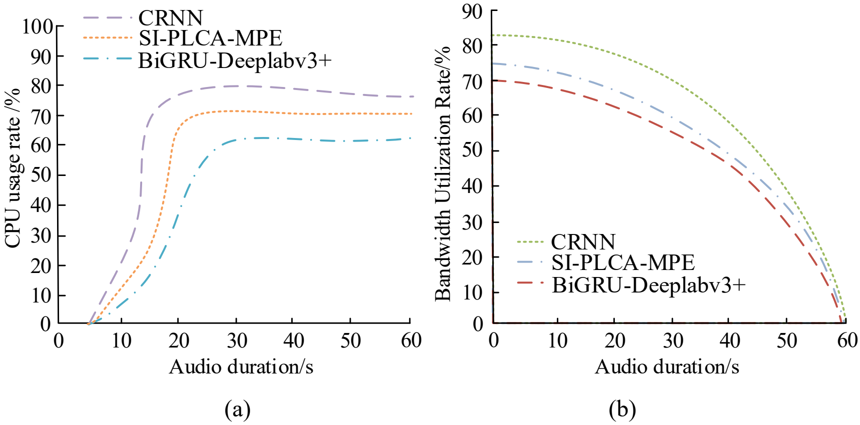
Figure 11 Multi-instrument polyphonic transcription accuracy results, (a) MMRR (b) note accuracy (see online version for colours)



Figures 11(a) and 11(b) show the trends in MMRR and note accuracy for the three models at different audio complexity levels. Figure 11(a) shows that as the audio complexity gradually increases from level 1 to level 4, the MMRR of the CRNN and SI-PLCA-MPE methods exhibit a gradual increase, reaching 88% and 89%, respectively. However, there are significant overall fluctuations and limited increases. In contrast, BiGRU-DeepLabv3+ demonstrates greater stability and robustness, with MMRR consistently maintaining above 90% and reaching 96.4% in high-complexity scenarios. This indicates that the model has stronger temporal alignment and note boundary reconstruction capabilities when processing complex audio segments with multiple tracks, harmonic overlaps, or intertwined rhythms. Figure 11(b) shows that the overall trend is basically consistent with MMRR, but the data is more distinctive. BiGRU-DeepLabv3+ achieves a note accuracy rate of over 93% at all levels of complexity, reaching 97.6% at the highest level of complexity. The traditional CRNN method achieves a recognition accuracy of 86% at low complexity, while SI-PLCA-MPE performs slightly better than CRNN, maintaining a lead of approximately 1.5%–2% across all levels. However, it still lags behind BiGRU-DeepLabv3+ by more than 5% in terms of performance. In summary, BiGRU-DeepLabv3+ outperforms other models in terms of structural capture and time series modelling. It effectively enhances the overall stability and recognition accuracy of multi-instrument transcription systems when faced with challenges such as high-frequency signal overlap, increased instrument variety, and complex rhythmic changes. The audio duration represents the continuous duration of the input audio segment, ranging from 10 s to 60 s, reflecting the changes in computational efficiency of the model at different audio lengths. A comparison of resource consumption among different methods during the transcription task is shown in Figure 12.

Figure 12(a) and Figure 12(b) show the changes in CPU usage and bandwidth utilisation of the three methods at different audio durations. In Figure 12(a), as the audio duration increases, the CPU usage of the three methods first rises rapidly and then stabilises. The CRNN method has the highest resource consumption. When the audio length reaches 20 seconds, its CPU usage stabilises at around 75%, while SI-PLCA-MPE remains slightly below this value, stabilising at around 70%. In contrast, BiGRU-DeepLabv3+ has more controllable resource consumption, with CPU usage stabilising at around 60%, demonstrating its excellent computational efficiency and lightweight advantages.

Figure 12 Comparison of resource consumption, (a) CPU usage rate (b) bandwidth utilisation rate (see online version for colours)



In Figure 12(b), the bandwidth utilisation rates of all three methods gradually decrease as the audio length increases. Among them, the bandwidth utilisation of CRNN and SI-PLCA-MPE exceeds 80% and 75%, respectively, at the initial stage, and after the audio duration reaches 40s, the bandwidth utilisation of both models drops rapidly to below 35%. In contrast, BiGRU-DeepLabv3+ maintains a much smoother decreasing trend throughout the process, indicating that the model's bandwidth consumption is more stable in long-duration audio processing. In addition, the model is still able to maintain 35% bandwidth utilisation at 50s, which is significantly better than the other two methods. It shows that the model is suitable for deployment in real-world scenarios with limited bandwidth or high transmission stability requirements.

Table 4 Comparison of transcriptional performance of models under different noise conditions

SNR	*Model*	*F1-score (%)*	*Pitch deviation (semitones)*	*Rhythm deviation (ms)*	*F1-score decrease (vs. 25 dB, %)*	*Inference delay (ms/frame)*
25 dB	CRNN	85.2	±0.29	22	/	50
	SI-PLCA-MPE	82.6	±0.34	27	/	48
	GRU-DeepLabv3+	90.4	±0.16	10	/	35
15 dB	CRNN	79.1	±0.41	31	−7.2%	52
	SI-PLCA-MPE	76.8	±0.45	36	−7.0%	50
	GRU-DeepLabv3+	89.1	±0.24	16	−1.3%	38
5 dB	CRNN	71.6	±0.58	47	−13.6%	55
	SI-PLCA-MPE	69.2	±0.61	51	−13.5%	53
	GRU-DeepLabv3+	86.4	±0.32	19	−4.0%	40

In order to further evaluate the robustness and generalisation ability of the proposed BiGRU-DeepLabv3+ transcription network in real acoustic environments, additional experiments were conducted using audio samples with different levels of background noise. Specifically, three SNR conditions were tested, namely 25 dB, 15 dB, and 5 dB, while keeping all other settings consistent. In addition to the F1 score indicator, pitch deviation and rhythm deviation were also introduced to quantify the accuracy of note frequency estimation and time alignment, and the degradation of F1 score relative to high SNR was calculated to evaluate noise sensitivity. Pitch deviation measures the degree of deviation between the output note frequency of the model and the true pitch. Rhythm deviation reflects the time deviation between the start and end times of the notes recognised by the model and the true annotations. The smaller the value, the more accurate the rhythm positioning. The results are shown in Table 4.

In Table 4, under the condition of 25 dB, the F1 score of GRU-DeepLabv3+ is 90.4%, which is 5.2% and 7.8% higher than CRNN and SI-PLCA-MPE, respectively. At the same time, the pitch deviation is controlled within ±0.16 semitones, and the rhythm deviation is only 10 ms, indicating that it has strong ability to recover musical score structures in high-quality audio. As the SNR decreases to 15 dB, the recognition performance of traditional models shows a significant decline. The F1 score of CRNN drops to 79.1%, the rhythm deviation increases to 31 ms, and the pitch deviation also expands to ±0.41 semitone. SI-PLCA-MPE also showed a similar trend. However, GRU-DeepLabv3+ only decreased by 1.3%, still maintaining an F1 score of 89.1%. The pitch deviation and rhythm deviation were ±0.24 semitones and 16 ms, respectively, indicating that the model can still effectively capture note boundaries and spectral structures under moderate noise interference. Under the condition of 5 dB, the recognition performance of CRNN and SI-PLCA-MPE decreased to 71.6% and 69.2%, respectively. Compared with clear scenes, F1 score decreased by more than 13%, pitch deviation reached ±0.58 to ±0.61 semitones, and rhythm error exceeded 47 ms, indicating severe noise interference in spectral feature extraction. However, the F1 score of GRU-DeepLabv3+ remained at 86.4%, only decreasing by 4.0%, with a pitch deviation of ±0.32 semitones and a rhythm deviation of 19 ms, demonstrating strong noise resistance and structural recovery ability. In addition, the inference delay of GRU-DeepLabv3+ remained at 35–40 ms/frame at all noise levels, significantly lower than that of CRNN and SI-PLCA-MPE, verifying its engineering adaptability in real-time automatic transcription tasks.

4 Discussion and interpretation

Aiming at the problems of low accuracy and sensitivity to noise interference of multi-instrument polyphony automatic transcription, a multi-instrument polyphony automatic transcription method based on BiGRU-DeepLabv3+ network was proposed. The time series modelling capability was enhanced by introducing the BiGRU module, while the network structure was lightened and the structural design of the null convolution and decoder was optimised. In addition, the audio was pre-processed using the front source separation module, which effectively enhanced the source separation and denoising capability. The experimental results indicated that the F1-score of the proposed method was 92.1% and 89.9% on the training and test sets, respectively, which was significantly better than the other two comparison models. In addition, the average note F1-score of the proposed model reached 83.65% and the average note accuracy was

85.4% in the four test sets with different instrument combinations. Significant improvement over the comparative approach was demonstrated by the improvement in SDR and SNR to 12.8 dB and 13.7 dB, respectively.

The model's great generalisation capacity for multi-source complicated instrument structures was demonstrated by the F1-score, which reached 84.5%, particularly when electric guitar and percussion instruments were combined. Compared with the automatic percussion transcription model based on the CNN-LSTM structure proposed by Cahyaningtyas et al. (2023), the BiGRU structure used in the study has more advantages in dealing with long-term dependencies and effectively reduces the possible gradient explosion problem. Park et al. (2023) constructed a transformer melody transcription model that performed well in modelling long sequences. However, its model parameters were large and inference time was long, making it unsuitable for deployment on resource-constrained devices. In contrast, the method proposed by the research achieved structural compression and computational optimisation while maintaining high transcription performance, demonstrating strong practicality.

In addition, the model also performed well in terms of MMRR and note accuracy, achieving 96.4% and 97.6% at the highest complexity level, respectively. Resource utilisation analysis indicated that BiGRU-DeepLabv3+ had good system overhead control capabilities while maintaining high accuracy. However, the study still has certain limitations. For example, the model still exhibits some performance fluctuations under extreme reverberation or low SNR audio conditions, and it has high requirements for training resources. Future work will explore lightweight structures further to improve real-time performance. Additionally, attention mechanisms will be introduced to enhance the model's ability to distinguish between different instrument features. These efforts will drive the development of multi-instrument automatic transcription systems toward higher accuracy and broader adaptability.

5 Conclusions

The study proposed a multi-instrument polyphonic transcription model that integrated BiGRU with an improved DeepLabv3+ structure, innovatively introducing semantic segmentation ideas into the note recognition task of audio spectrograms. By enhancing the model's temporal modelling capabilities through BiGRU and combining it with the ASPP structure to achieve multi-scale semantic decoding, the accuracy of note recognition in polyphonic environments was significantly improved. The experiments showed that the proposed model outperformed the comparison model in key accuracy, precision, and average precision indexes, particularly in accuracy and recognition recall rates. There was a significant 6.8% improvement in note-level precision compared to the traditional model and a more than 23.5% improvement in inference efficiency. Meanwhile, the model also had better resource utilisation and reasoning efficiency, showing good engineering practicability and scalability. The research provides new ideas in the field of music information processing and theoretical and methodological support for the practical deployment of real-time, multi-instrument transcription systems. This lays the foundation for subsequent applications, such as intelligent music creation, music analysis, and human-computer interaction.

Declarations

All authors declare that they have no conflicts of interest.

References

Cahyaningtyas, Z.A., Purwitasari, D. and Fatichah, C. (2023) 'Deep learning approaches for automatic drum transcription', *EMITTER Int. J. Eng. Technol.*, Vol. 11, No. 1, pp.21–34, DOI: 10.24003/emitter.v11i1.764.

Chen, H., Qin, Y., Liu, X., Wang, H. and Zhao, J. (2024) 'An improved DeepLabv3+ lightweight network for remote-sensing image semantic segmentation', *Complex Intell. Syst.*, Vol. 10, No. 2, pp.2839–2849, DOI: 10.1007/s40747-023-01304-z.

Edwards, D., Dixon, S. and Benetos, E. (2023) 'Pijama: piano jazz with automatic MIDI annotations', *Trans. Int. Soc. Music Inf. Retrieval*, Vol. 6, No. 1, pp.89–102, DOI: 10.5334/tismir.162.

Guo, R. and Zhu, Y. (2025) 'Research on the recognition of piano-playing notes by a music transcription algorithm', *J. Adv. Comput. Intell. Intell. Inform.*, Vol. 29, No. 1, pp.152–157, DOI: 10.20965/jaciii.2025.p0152.

Ji, J., Li, S., Liao, X. and Zhang, F. (2022) 'Semantic segmentation based on spatial pyramid pooling and multilayer feature fusion', *IEEE Trans. Cogn. Dev. Syst.*, Vol. 15, No. 3, pp.1524–1535, DOI: 10.1109/TCDS.2022.3225200.

Lee, D. and Jeong, D. (2023) 'Reducing latency of neural automatic piano transcription models', *J. Acoust. Soc. Korea*, Vol. 42, No. 2, pp.102–111, DOI: 10.7776/ASK.2023.42.2.102.

Lee, J. and Lee, W.G. (2024) 'Direct visualization of bass guitar frequency patterns and their fret fingerings via combined fast and short-time Fourier transforms', *J. Vib. Eng. Technol.*, Vol. 12, No. 7, pp.7419–7428, DOI: 10.1007/s42417-024-01303-5.

Li, X., Yan, Y., Soraghan, J., Wang, Z. and Ren, J. (2023) 'A music cognition-guided framework for multi-pitch estimation', *Cogn. Comput.*, Vol. 15, No. 1, pp.23–35, DOI: 10.1007/s12559-022-10031-5.

Luo, H., Cao, Z., Zhang, X., Li, C. and Kong, D. (2022) 'Combining different forms of statistical energy analysis to predict vibrations in a steel box girder comprising periodic stiffening ribs', *Steel Compos. Struct.*, Vol. 45, No. 1, pp.119–131, DOI: 10.12989/scs.2022.45.1.119.

Mohamed, B. and Yassine, B.A. (2023) 'Enhanced video temporal segmentation using a Siamese network with multimodal features', *Signal Image Video Process.*, Vol. 17, No. 8, pp.4295–4303, DOI: 10.1007/s11760-023-02662-4.

Park, J., Choi, K., Oh, S., Kim, L. and Park, J. (2023) 'Note-level singing melody transcription with transformers', *Intell. Data Anal.*, Vol. 27, No. 6, pp.1853–1871, DOI: 10.3233/IDA-227077.

Peng, H., Yu, Y. and Yu, S. (2023) 'Re-thinking the effectiveness of batch normalization and beyond', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 46, No. 1, pp.465–478, DOI: 10.1109/TPAMI.2023.3319005.

Peng, L. (2023) 'Piano players' intonation and training using deep learning and MobileNet architecture', *Mob. Netw. Appl.*, Vol. 28, No. 6, pp.2182–2190, DOI: 10.1007/s11036-023-02175-x.

Preethi, P. and Mamatha, H.R. (2023) 'Region-based convolutional neural network for segmenting text in epigraphical images', *Artif. Intell. Appl.*, Vol. 1, No. 2, pp.119–127, DOI: 10.47852/bonviewAIA2202293.

Simonetta, F., Avanzini, F. and Ntalampiras, S. (2022) 'A perceptual measure for evaluating the resynthesis of automatic music transcriptions', *Multimed. Tools Appl.*, Vol. 81, No. 22, pp.32371–32391, DOI: 10.1007/s11042-022-12476-0.

Spoorthy, V. and Koolagudi, S.G. (2024) ‘Polyphonic sound event detection using Mel-pseudo constant Q-transform and deep neural network’, *IETE J. Res.*, Vol. 70, No. 5, pp.5031–5043, DOI: 10.1080/03772063.2023.2253768.

Suresh Kumar, S.A. and Rajan, R. (2023) ‘Transformer-based automatic music mood classification using multi-modal framework’, *J. Comput. Sci. Technol.*, Vol. 23, No. 1, pp.18–34, DOI: 10.24215/16666038.23.e02.

Talwar, S., Barbero, F.M., Calce, R.P. and Collignon, O. (2023) ‘Automatic brain categorization of discrete auditory emotion expressions’, *Brain Topogr.*, Vol. 36, No. 6, pp.854–869, DOI: 10.1007/s10548-023-00983-8.

Todjro, M. and Mensah, Y. (2023) ‘Convolution product for Hilbert C*-module valued maps’, *Sahand Commun. Math. Anal.*, Vol. 20, No. 3, pp.19–31, DOI: 10.22130/scma.2022.557582.1145.

Velazquez Lopez, O., Oropeza Rodriguez, J.L. and Suarez Guerra, S. (2022) ‘Application of auditory filter-banks in polyphonic music transcription’, *Comput. Sist.*, Vol. 26, No. 4, pp.1421–1428, DOI: 10.13053/CyS-26-4-4271.

Wang, P. and Dai, N. (2025) ‘Processing piano audio: research on an automatic transcription model for sound signals’, *J. Meas. Eng.*, Vol. 13, No. 1, pp.130–139, DOI: 10.21595/jme.2024.24345.

Wang, W., Li, J., Li, Y. and Xing, X. (2024) ‘Style-conditioned music generation with transformer-GANs’, *Front. Inf. Technol. Electron. Eng.*, Vol. 25, No. 1, pp.106–120, DOI: 10.1631/FITEE.2300359.

Wang, Y. (2023) ‘Piano automatic transcription based on transformer’, *J. Intell. Fuzzy Syst.*, Vol. 45, No. 5, pp.8441–8448, DOI: 10.3233/JIFS-233653.

Wang, Y. and Dong, Y. (2024) ‘A semantic segmentation method of remote sensing image based on feature fusion and attention mechanism’, *J. Inf. Process. Syst.*, Vol. 20, No. 5, pp.640–653, DOI: 10.3745/JIPS.01.0108.

Wang, Y., Feng, S., Wang, B. and Ouyang, J. (2023) ‘Deep transition network with gating mechanism for multivariate time series forecasting’, *Appl. Intell.*, Vol. 53, No. 20, pp.24346–24359, DOI: 10.1007/s10489-023-04503-w.

Wu, P., Guo, R., Tong, X., Su, S., Zuo, Z., Sun, B. and Wei, J. (2022) ‘Link-RGBD: cross-guided feature fusion network for RGBD semantic segmentation’, *IEEE Sens. J.*, Vol. 22, No. 24, pp.24161–24175, DOI: 10.1109/JSEN.2022.3218601.

Xu, G., Guo, W. and Wang, Y. (2023) ‘Subject-independent EEG emotion recognition with hybrid spatio-temporal GRU-Conv architecture’, *Med. Biol. Eng. Comput.*, Vol. 61, No. 1, pp.61–73, DOI: 10.1007/s11517-022-02686-x.

Xu, M. and He, J. (2023) ‘Seispy: Python module for batch calculation and postprocessing of receiver functions’, *Seismol. Soc. Am.*, Vol. 94, No. 2, pp.935–943, DOI: 10.1785/0220220288.

Yang, D., Du, Y., Yao, H. and Bao, L. (2022) ‘Image semantic segmentation with hierarchical feature fusion based on deep neural network’, *Connect. Sci.*, Vol. 34, No. 1, pp.1772–1784, DOI: 10.1080/09540091.2022.2082384.

Yi, X., Zhang, S. and Zhou, Y. (2024) ‘Efficient 2-D MUSIC algorithm for super-resolution moving target tracking based on an FMCW radar’, *Geod. Geodyn.*, Vol. 15, No. 5, pp.504–515, DOI: 10.1016/j.geog.2024.01.008.