Yao *et al.*
EURASIP Journal on Advances in Signal Processing (2025) 2025:39
https://doi.org/10.1186/s13634-025-01249-0

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

TFSWA-ResUNet: music source separation with time–frequency sequence and shifted window attention-based ResUNet

Zhenyu Yao[1,2], Yuping Su[1,2*], Honghong Yang[2,3], Yumei Zhang[1,2,3] and Xiaojun Wu[1,2,3]

*Correspondence:
ypsu@snnu.edu.cn

[1] School of Artificial Intelligence and Computer Science, Shaanxi Normal University, Xi'an, China
[2] Key Laboratory of Intelligent Computing and Service Technology for Folk Song, Ministry of Culture and Tourism, Xi'an, China
[3] Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, China

Abstract

CNN-based UNet is a widely employed network architecture for music source separation (MSS). Meanwhile, spectrogram features of music audio, which consist of both time and frequency information, are commonly used as inputs for MSS tasks. CNN-based UNet models in the spectrogram domain still have the limitation that the global and local correlations of spectrograms have not been explored efficiently. In this paper, we propose a novel ResUNet architecture named TFSWA-ResUNet, in which a temporal-frequency and shifted window attention (TFSWA)-based module is designed as UNet's bottleneck block. In the proposed TFSWA block, the time sequence attention (TSA) block and frequency sequence attention (FSA) block are used to capture the global correlations of music spectrogram features within time and frequency sequence, respectively. To further capture the local correlations of spectrogram features, a shifted window attention-based Swin transformer is also introduced into the TFSWA module, which computes self-attention within local non-overlapping windows and captures correlations from both the temporal and frequency dimensions. Experimental results on the MUSDB18 dataset indicate that the proposed TFSWA-ResUNet model can achieve significant separation performance with a relatively small number of parameters, which demonstrates that our approach offers a good tradeoff between performance and computational cost, making it a feasible and widely adoptable solution.

Keywords: Music source separation, ResUNet, Time sequence attention, Frequency sequence attention, Shifted window attention

1 Introduction

Music source separation (MSS) is a task of decomposing audio mixtures into several distinct musical tracks, including *vocals*, *bass*, *drums*, *etc.* [2]. It is a key research topic for music information retrieval (MIR) [7] and has found wide applications in melody extraction [27], music transcription [19], beat tracking [45], and so on.

With the rapid development of deep learning techniques, deep neural network-based approaches have shown impressive progress in recent years [11, 14, 18, 25, 34]. Most approaches in MSS operate on the spectrograms generated by the short-time Fourier transform (STFT). To recover the signal, the inverse short-time Fourier transform

(ISTFT) is applied to the predicted spectrogram. In MSS studies, most works are either convolutional neural network (CNN)-based models with a U-shaped structure [5, 15, 16, 21] or recurrent neural network (RNN)-based models [26, 37]. The typical U-shaped network, known as UNet [32], consists of a symmetric encoder-decoder architecture with skip connections. The encoder uses a series of convolutional layers followed by consecutive down-sampling layers to extract deep features with large receptive fields. The decoder then up-samples these deep features back to the input resolution to reconstruct the target source.

In this paper, we focus on the field of CNN-based MSS methods. Although CNN deep learning methods have become increasingly popular in recent years, they still face several challenges and limitations when applied to MSS tasks. Firstly, due to the inherent locality of the convolution operation, CNN-based approaches are unable to capture global and long-range information interactions [3]. Another challenge is that traditional CNN models do not consider the different degrees of influence from features at varying spatial positions or time intervals, and treat all features as equally important. Fortunately, the emergence of the transformer-based attention mechanism [40] has largely addressed these issues. The attention mechanism allows the model to perform weighted calculations on different parts of the input, so the network can focus more on important information and ignore irrelevant details [40]. Besides, the attention mechanism enables the model to capture dependencies between elements regardless of their distance, enhancing the understanding of long-range dependencies. To incorporate attention mechanisms into image processing field, the vision transformer (ViT) is proposed [6]. The core idea of ViT is to model the relationships between image patches using the transformer's attention mechanism. However, ViT employs global self-attention, which computes relationships between a token and all other tokens. This global computation results in quadratic complexity relative to the number of tokens, making it unsuitable for many vision problems that require a large number of tokens to represent high-resolution images. To address this problem, an improved ViT architecture called Swin transformer [22] is proposed as a vision backbone. Based on the shifted window mechanism, the Swin transformer achieves greater efficiency by limiting self-attention computation to non-overlapping local windows, while also allowing for cross-window connections.

Motivated by existing works, a novel spectrogram-domain method for the MSS task named TFSWA-ResUNet is proposed in this paper. In the proposed method, a temporal-frequency and shifted window attention (TFSWA)-based bottleneck block is designed for the basic ResUNet architecture and it can efficiently capture the global and local correlations of music spectrogram features. The contributions of this paper can be summarized as follows:

(1) We propose a novel attention-based block called TFSWA, which consists of a time sequence attention (TSA) block, a frequency sequence attention (FSA) block, followed by a residual branch consisting of a Swin transformer block. In particular, the TSA block computes multi-head attention for the spectrogram feature maps along the time axis, and global attention is computed within the time sequence for each frequency bin. FSA blocks apply multi-head attention along the frequency axis to model the global dependencies of spectrogram features at the same time frame. The

Swin transformer block consists of two consecutive modules which perform multi-head self-attention on local windows and shifted windows, respectively.

(2) We construct a TFSWA-based ResUNet architecture called TFSWA-ResUNet for MSS which predicts the source magnitude spectrogram directly. The proposed TFSWA-ResUNet consists of five encoder blocks, five decoder blocks and four TFSWA blocks serving as intermediate block. Each encoder and decoder block is composed of a light 2-layer Conv block and a down-sampling/up-sampling layer, which reduces the parameter size compared to the conventional UNet structure. Besides, the self-attention computation of Swin transformer module in TFSWA block is performed within local non-overlapping windows, which can further reduce the computation cost. Simulation results on the MUSDB18 dataset [31] show that the proposed model is able to achieve good performance with a relatively small number of parameters.

(3) For the TFSWA block, TSA captures the long-range contextual information for each time sequence, and FSA captures the global frequency correlation for each frequency sequence. Finally, the shifted window attention in the Swin transformer captures the local spectrogram features from both time and frequency dimensions. Experimental results demonstrate that the combination of these attention blocks within the TFSWA block significantly enhances the representation capability of the proposed model, leading to an average SDR improvement of 0.34 dB across four music sources.

The remainder of this paper is organized as follows. Related works are provided in Section 2. Section 3 illustrates the details of the proposed TFSWA-ResUNet model. The detailed experimental settings are described in Section 4. The experimental results and analysis are presented in Section 5. Finally, Section 6 concludes this paper and discusses the future research direction.

2 Related works

MSS models can typically be divided into three categories: time-domain approaches, spectrogram-domain approaches and hybrid approaches. Time-domain approaches directly model the raw audio waveform to separate multiple sources [9, 25], while spectrogram-domain methods operate on spectrogram to extract target sources [17, 21]. In recent years, hybrid approaches are also proposed for MSS tasks that attempted to combine time-domain and spectrogram-domain methods. These methods are discussed in detail in the following subsections.

2.1 Spectrogram-domain methods

Spectrogram-domain-based methods operate on the mixture spectrogram, which is estimated from the waveform using the STFT. The spectrogram of the individual sources are approximated from the mixture spectrogram either by directly predicting the spectrogram representation of each source [28, 39], or estimating a mask for each source to recover the individual sources [13, 17, 21].

In particular, authors in [13] propose to apply deep UNet convolutional network to the music source separation. The model outputs a soft mask that is multiplied with the

mixture spectrogram to obtain the estimate spectrogram of vocal and instrumental components. In order to get larger receptive field and enable the local and global feature information to be modeled simultaneously within a single convolution layer, authors in [39] proposed a novel CNN model called D3Net for MSS which combines the multi-dilated convolution with DenseNet architecture to model multi-resolution spectrogram features. The outputs of D3Net network are used to calculate the multichannel Wiener filter (MWF) to obtain the final separations, as commonly performed in frequency-domain audio source separation methods [28]. Conventional spectrogram-based methods usually reuse the mixture phase when estimating the sources and this may limit the separation performance. To address this issue, authors in [4] propose to use Complex as Channel (CaC) framework based on complex-valued spectrogram estimation, which considering real and imaginary parts of a spectrogram as separate real-valued channels. As an alternative solution, authors in [17] propose to estimate the complex ideal ratio masks (cIRMs) by decoupling the spectrogram magnitude and phase [43]. Besides, they also extend the separation method to effectively allow the magnitude of the mask to be larger than 1. In reference [21], a channel-wise subband phase-aware ResUNet (CWS-PResUNet) is proposed which further utilizes a channel-wise subband spectrogram feature to limit unnecessary global weights sharing on the spectrogram and reduce computational resource consumption. Recently, authors in [26] propose a band-split RNN (BSRNN) which splits the spectrogram of the mixture into subbands and utilizes BLSTM to perform interleaved band-level and sequence-level modeling. Motivated by BSRNN method, a novel frequency-domain approach named BS-RoFormer which is based on a band-split RoPE transformer architecture is proposed in [24]. BS-RoFormer employs a band-split module to project the input complex spectrogram into subband-level representations, and then arranges a stack of hierarchical Transformers to model the inner-band as well as inter-band sequences for multi-band mask estimation. It achieves 9.80 dB of average SDR on the MUSDB18HQ dataset. As an extension of BS-RoFormer, Mel-RoFormer adopts the Mel-band scheme that maps the frequency bins into overlapped subbands according to the mel scale, and it achieves better performance than BS-RoFormer in the MSS tasks [42].

Spectrogram-domain methods can fully utilize the rich information contained in the spectrogram and obtain significant performance, so they are widely used in MSS tasks.

2.2 Time-domain methods

Most spectrogram-domain methods only consider the spectrogram magnitude estimation, and the phase information of the target source directly adopts that of the mixed source. This may lead to inaccuracy of the target source's spectral phase information [13, 39]. Although some spectrogram-domain methods have been proposed to estimate the phase information by decoupling the magnitude and phase for estimating complex ideal ratio masks [17, 21], the estimation becomes complicated and increases the computation cost. Therefore, there are some works investigating source separation in the time domain to directly model the phase information.

For example, authors in [36] propose the Wave-UNet for music source separation, an adaptation of the UNet [13] to the one-dimensional time domain. The Wave-UNet consists of a 12-layer encoder to down-sample the feature maps and 12-layer decoder

to up-sample the feature maps respectively, and each decoder layer access to the corresponding encoder layer via skip connections. Experiments show that it outperforms the state-of-the-art spectrogram-based UNet architecture [13] at the time. In reference [23], authors propose a novel convolutional Wavenet for MSS with audio waveform as input and output. The new Wavenet is an adaptation of Wavenet [29] that turns the original dilated causal convolution Wavenet into a non-causal model. With the non-causal adaptation in the novel Wavenet, the model is able to predict a target field instead of one sample at a time. Experiment results show that this novel Wavenet performs comparably to Wave-UNet [36]. Similarly, authors in [9] propose a novel waveform-to-waveform model called Demucs, with a combination of a UNet structure and bidirectional LSTM. The UNet architecture consists of a convolutional encoder and a decoder based on wide transposed convolutions with large strides. To increase the number of channels exponentially with depth, a bidirectional LSTM is also introduced between the encoder and the decoder. Experiments show that, with proper data augmentation, Demucs can achieve better performance with 6.3 SDR on average.

Although time-domain-based MSS methods can naturally take into account phase information, their performance may be suboptimal due to the high dimensionality of the time-domain signal and the difficulty of sufficient feature extraction.

2.3 Hybrid domain methods

To fully integrate the advantages of wave-domain and spectrogram-domain methods in music source separation, some research proposes employing hybrid domain methods to leverage the features of both domains and achieve better separation performance.

In particular, hybrid Demucs [8] expands the Demucs framework [9] to perform MSS in a hybrid waveform/spectrogram domain. It proposes a dual UNet structure to model the time and spectral-domain separately and employ a shared encoder to integrate two outputs. The temporal branch takes the input waveform, and the spectral branch takes the spectrogram obtained from an STFT. The model's final prediction is obtained by summing the inverse STFT-processed output of the spectral branch with the output from the temporal branch. The model can freely choose the most convenient representation for different parts of the signal and can share information between the two representations without restrictions. Motivated by hybrid Demucs [8], a fully hybrid waveform and spectrogram generative adversarial network is presented for MSS [44]. In particular, the generator is innovated by the structure of Hybrid Demucs. To further utilize long-range contextual information of music, hybrid transformer Demucs (HT Demucs) [33] replaces the innermost convolutional layers of the original Hybrid Demucs [8] by cross-domain transformer layers to capture the sequence correlation. The transformer layers use self-attention within one domain and cross-attention across domains to process spectral and temporal information. Experiments indicate that HT Demucs outperforms Hybrid Demucs when utilizing additional training songs. Besides, CDE-HTCN model [12] combines the complexed spectrogram-domain feature with time-domain feature using a cross-domain encoder (CDE) to leverage the advantages of both domains. It also employs a powerful separator named hierarchic temporal convolutional network (HTCN) for multiple music sources separation. The experimental results show that CDE-HTCN outperforms the state-of-the-art methods at the time.

Hybrid domain methods can improve separation performance to a certain extent, but the limitation is the increased complexity of the cross-domain encoder/decoder, requiring careful alignment of the temporal and spectral signals through well-shaped convolutions and proper parameter setting.

3 The proposed TFSWA-ResUNet model

3.1 Architecture overview

Figure 1 shows the overall structure of the proposed TFSWA-ResUNet model, which is fundamentally a ResUNet architecture consisting of five encoder blocks, five decoder blocks and four TFSWA blocks serving as the bottleneck blocks. For the proposed model, each channel of the input stereo mixture waveform is first transformed into spectrogram magnitude via short-time Fourier transform. Then, each full-band spectrogram is split into several subbands along the frequency dimension. Finally, all the subbands of the stereo signal are stacked together to obtain an multichannel feature maps. In this way, each subband is treated as a channel, allowing the model to process different channels independently and capture the inherent patterns of different frequency ranges. On the decoder side, the last decoder block outputs the target source spectrogram magnitude, and these features are flattened and merged along the channel dimension. The merged spectrogram magnitude are then combined with the phase of the mixture to recover the target waveform through inverse short-time Fourier transform (ISTFT).

3.2 Encoder and decoder of the proposed model

The proposed model is constructed based on a basic ResUNet structure. As shown in Fig. 1, the basic network is constructed with five encoder and decoder blocks. For a stereo mixture waveform for separation, we first obtain its magnitude spectrogram

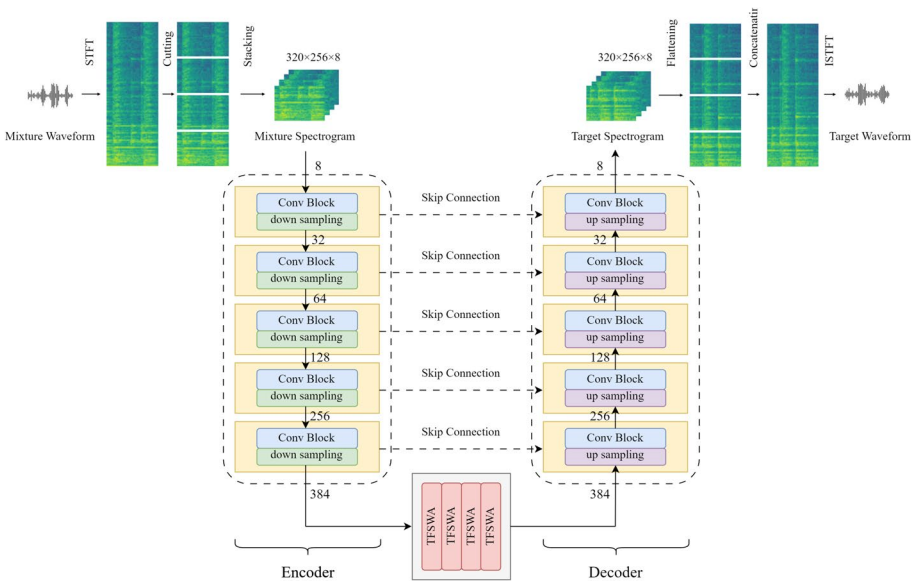

Fig. 1 Overall architecture of the proposed TFSWA-ResUNet model, which consists of five encoder blocks, five decoder blocks, and four TFSWA blocks as bottleneck for modeling the temporal and frequency dimension correlation within feature maps. The numbers 8, 32, 64, 128, 256 and 384 in encoder and decoder blocks denote the corresponding number of feature maps

through STFT. It has been shown that different subbands may contain different local patterns [20], so we split the full-band spectrogram of each track into four subbands. As a result, eight subbands are obtained for the stereo mixture signal. These subbands are then stacked together to obtain an eight-channel feature maps. In our model, the stacked input spectrogram is of size $320 \times 256 \times 8$, in which 320 is the number of time frames, 256 is the number of frequency bins and 8 is the number of channels.

For the encoder of the proposed model, each encoder block consists of a Conv block and a down-sampling block. The Conv block contains four residual convolutional modules (RCM) as shown in Fig. 2. Each RCM includes two convolutional layers with a kernel size of 3×3 and a stride size of 1×1. A batch normalization and a GELU activation [10] is applied respectively before two convolutional layers. A shortcut connection is added between the input and the output of a RCM through a 1×1 convolution. The down-sampling module of encoder blocks consists of a 2×2 average pooling with a stride of 2. For each encoder block, the Conv block keeps the feature size fixed but doubles the number of feature maps except the first and last blocks, and each down-sampling module reduces the feature map size by half. Multiple encoder blocks allow the model to learn multi-scale features with different resolutions.

The blocks in the decoder are symmetric to those in the encoder. Each decoder block consists of a Conv block and an up-sampling layer. The Conv block has the same structure as those in the encoder blocks, but it reduces the number of channels in the feature maps while maintaining the feature size. The up-sampling layer employs bilinear interpolation [30] to up-sample the feature maps by a factor of two. The output feature maps of the Conv blocks on the encoder side are concatenated with the corresponding feature maps at the same resolution on the decoder side. This allows for more layer interactions and the reuse of previously computed features without loss of information from the successive down-sampling operations. The decoder finally outputs the target source magnitude spectrogram with the same size as the model's input. Based on the output spectrogram and the mixture phase information, the target waveform is recovered through ISTFT. The number of channels in the input and output feature maps for each encoder and decoder block is illustrated in Fig. 1.

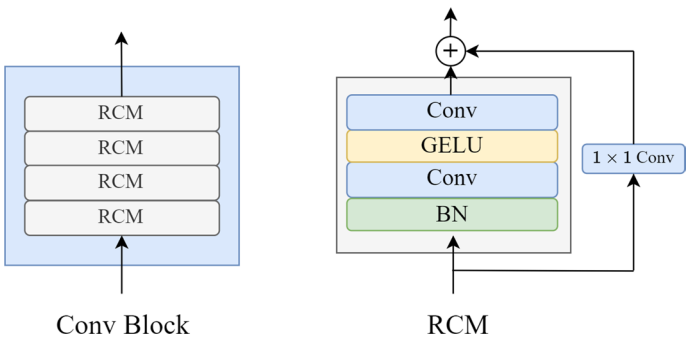

Fig. 2 Details of Conv Block in the encoder and decoder

3.3 TFSWA module: intermediate block of the proposed model

To further increase the representation ability of the basic residual UNet, a time–frequency sequence and shifted window attention (TFSWA)-based module is proposed as intermediate blocks between the encoder and decoder. The intermediate block consists of four TFSWA modules, each of which includes a time sequence attention (TSA) block, a frequency sequence attention (FSA) block, followed by a residual branch consisting of a Swin transformer (a transformer using shifted window attention) block, as shown in Fig. 3(a). Specifically, the TSA block is employed to capture the global temporal correlation of spectrogram features. The FSA is used to identify the global frequency dependency of spectrogram features. Lastly, the shifted window attention within the Swin Transformer is utilized to capture the local temporal-frequency correlations of the spectrogram.

TSA and FSA Blocks: In the TFSWA module, the TSA and FSA blocks have the same structure as illustrated in Fig. 3(b). Each block consists of a LayerNorm (LN) [1], a multi-head self-attention (MSA) layer, followed by a two-layer MLP with a GELU activation function in between, and a residual connection is applied between the input and output of the MLP layer. The structure of multi-head self-attention mechanism is illustrated in Fig. 4. It linearly projects the queries, keys, and values h times respectively, and then performs scaled dot-product self-attention in parallel. The outputs from each head are concatenated and then projected to generate the final output. The scaled dot-product self-attention in the MSA for each head is computed as [40]:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{S\times d}$ indicate the query, key and value matrices, respectively; d is the dimension of the query and key, and S represents the length of the time sequence or frequency sequence to which self-attention is applied.

Although the TSA and FSA blocks share the same structure, they perform MSA on the spectrogram features from different dimensions. As shown in Fig. 5, the TSA block

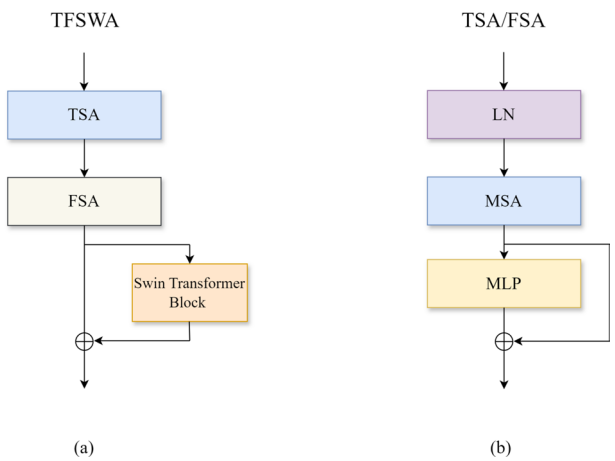

Fig. 3 **a** Structure of the TFSWA module. **b** Details of time sequence attention (TSA) block and frequency sequence attention (FSA) block, they share the same structure except that they perform multi-head self-attention (MSA) from different dimensions of feature maps

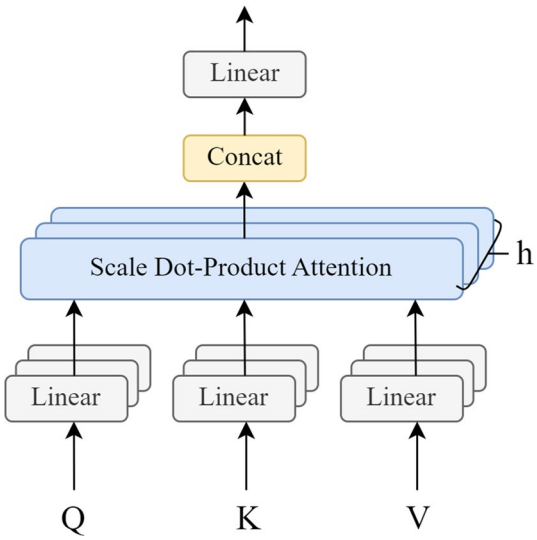

Fig. 4 Structure of multi-head self-attention (MSA) which consists of several attention layers running in parallel

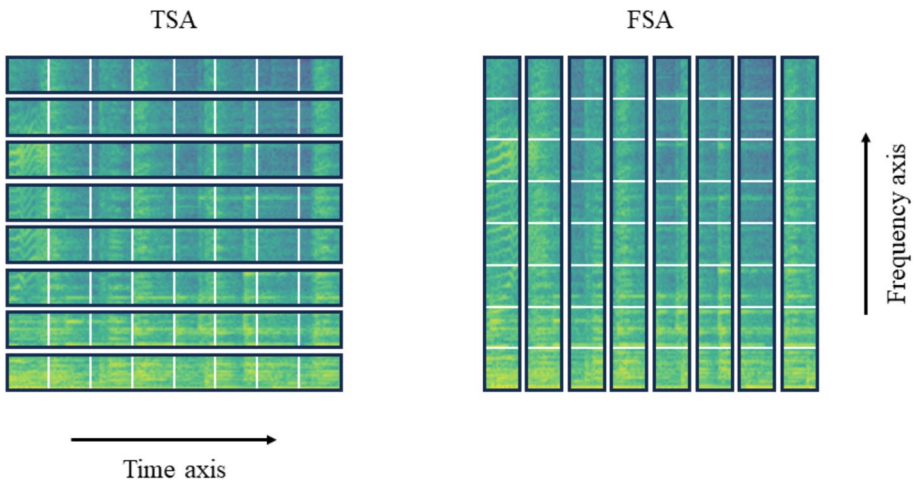

Fig. 5 TSA performs MSA along the time axis on input feature maps and computes self-attention within each time sequence that belongs to the same frequency bin. FSA performs MSA along the frequency axis on input feature maps and computes self-attention within each frequency sequence that belongs to the same time frame

computes multi-head self-attention for the input feature maps along the time axis. Self-attention is computed within the time sequence for each frequency bin and is applied to multiple frequency bins in parallel. The TSA mechanism captures the global temporal correlation of music spectrogram features for each frequency bin. The FSA applies multi-head self-attention along the frequency axis to model the dependencies of features at the same time frame, and self-attention is applied to multiple time frames in parallel.

To illustrate how the FSA(or TSA) block can capture the global information of input feature sequences, we visualize the attention weights of MSA module of the FSA in the last TFSWA block, as shown in Fig 6. In the FSA/TSA block, 4-head MSA is employed. In Fig 6, each heatmap displays the attention weights of Query positions on Key positions

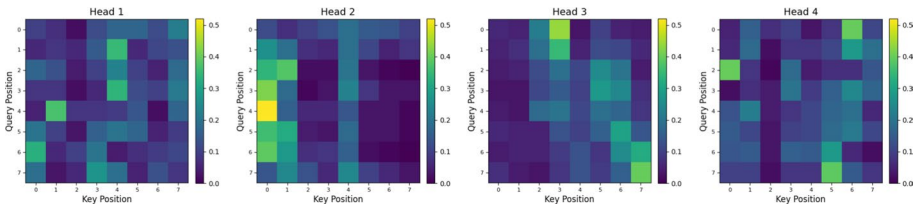

Fig. 6 Attention weight visualization of FSA blocks

for an input sequence with length of 8. From the figure, we observe distinct patterns across different attention heads, demonstrating their ability to capture varying contextual information. What's more, heads 1, 2, and 4 clearly demonstrate the capture of global information, as evidenced by high attention weights being extensively distributed across the entire matrix, rather than being confined to the vicinity of the diagonal. This suggests that when computing its output, a position attends to numerous other elements in the sequence, including those at a considerable distance. By analogy, TSA blocks, which operate similarly along the time axis, can also capture the global information.

Compared to a normal convolution layer that can model only local patterns, TSA and FSA conduct self-attention globally within temporal and frequency sequences, making them more capable of modeling the temporal correlation in beat patterns of music audio, as well as the frequency correlation in chorus and harmony, both of which are crucial for MSS tasks.

Shifted Window Attention in Swin Transformer Blocks: To further enhance the modeling capability of the TFSWA block while maintaining a relatively low computation cost, we introduce the Swin transformer block [22] in TFSWA through residual connection. Unlike TSA and FSA which model global temporal or frequency dependencies, the Swin transformer blocks employ window self-attention and shifted window self-attention to capture local temporal-frequency correlations in the spectrogram. The combination of global and local attention mechanisms enhances the representational ability of TFSWA.

The Swin transformer is constructed by replacing the traditional self-attention module in a Transformer block with one that computes self-attention within local windows, while keeping all other layers unchanged [22]. Window-based local attention can bring greater efficiency and lower computation complexity than global attention. Since window-based self-attention lacks connections across windows, the Swin transformer further incorporates a shifted window partitioning strategy within its blocks. This strategy switches between two different partitioning setups across successive Swin transformer blocks [22].

The structure of two consecutive Swin transformer blocks is illustrated in Fig. 7. It can be seen that the two blocks are essentially residual networks. They have the same structure expect for the multi-head self-attention (MSA) layer. The first block performs multi-head self-attention by dividing the input features into non-overlapping windows and performing attention computation within each window (i.e., W-MSA). To introduce connections between neighboring windows in the previous layer, the second consecutive module performs MSA by adopting a windowing configuration that is shifted from that of the preceding layer (i.e., SW-MSA), thus providing connections among them. The

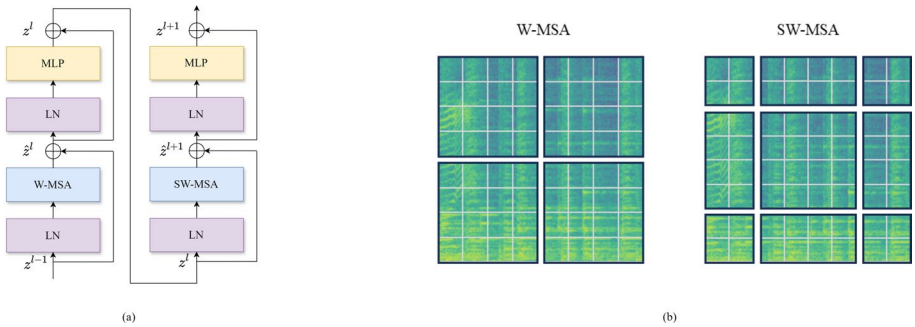

Fig. 7 **a** Structure of Swin transformer blocks in TFSWA module. **b** Window partition strategy of W-MSA and SW-MSA. The self-attention computation in the windows partitioned by SW-MSA crosses the boundaries of windows partitioned by W-MSA, providing connections between neighboring windows in the previous layer

window partition strategy of W-MSA and SW-MSA of a 8×8 feature map is illustrated in Fig. 7. Based on such window partitioning mechanism, continuous Swin transformer blocks can be formulated as:

$$\hat{z}^{l} = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \tag{2}$$

$$z^{l} = \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l}, \tag{3}$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^{l})) + z^{l}, \tag{4}$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}, \tag{5}$$

where $\hat{z}^{l}$ and z^{l} represent the outputs of the W-MSA module and the MLP module of the l^{th} block, respectively. $\hat{z}^{l+1}$ and z^{l+1} have the similar meanings.

In each TFSWA module, two consecutive Swin transformer blocks are employed, and the window size for both W-MSA and SW-MSA is 5×4. The W-MSA and SW-MSA layers compute multi-head attention within local windows that encompass both temporal and frequency dimensions. This effectively supplements the previous layers, TSA and FSA, which primarily capture correlations from either the temporal or frequency dimensions.

Overall, the TFSWA intermediate block captures correlations in spectrogram features through global time sequence attention in the TSA block, global frequency sequence attention in the FSA block, and local time–frequency attention in the Swin transformer block. It computes correlations of feature maps in both global and local manners, significantly enhancing the modeling power of the proposed model.

4 Experiment configurations

4.1 Dataset

We conduct a series of experiments to demonstrate the proposed method on the MUSDB18 dataset [31]. The dataset includes separate *vocals*, *drums*, *bass* and *other* instruments. It consists of a train set with 100 songs and a test set with 50 songs. All songs are in stereo format and sampled at 44.1 kHz.

4.2 Evaluation metrics

We mainly adopt the source-to-distortion ratio (SDR) [41] as objective evaluation metric to evaluate the proposed model and comparison methods. In the ablation experiments, we also added the source-to-interference ratio (SIR) and source-to-artifacts ratio (SAR) metric to demonstrate the effectiveness of the proposed model [41]. Given an estimate of a source s_i composed of the true source s_{target}, along with three error terms, interference e_{interf}, noise e_{noise}, and artifacts e_{artif}, the SDR, SIR and SAR can be defined as follows:

$$\mathrm{SDR} = 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}} + e_{noise} + e_{\mathrm{artif}}\|^2} \tag{6}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}\|^2} \tag{7}$$

$$\mathrm{SAR} = 10\log_{10}\frac{\|s_{\mathrm{target}} + e_{interf} + e_{\mathrm{noise}}\|^2}{\|e_{\mathrm{artif}}\|^2} \tag{8}$$

The SDR, SIR and SAR are calculated using the *museval* package [38]. Higher SDR, SIR, and SAR values all indicate better separation results, and vice versa. In addition, we also conduct a subjective evaluation using the mean opinion score (MOS) as the evaluation metric. The evaluation scores range from 1 to 5, with 5 being the highest score.

4.3 Experimental setup

For training, we split the audio into segments of 3 s each. For data augmentation, we employ mix-audio data augmentation proposed in [17], which randomly mixes two 3-second segments from a same source as a new 3-second segment for training. The input audio segments are transformed into spectrogram using STFT with a window size of 2048 samples and a hop size of 441 samples.

We train a dedicated model for each source since it is beneficial. We set batch size to 16 and we train for 30 epochs for *vocals* and *bass*. For *drums* and *other* instruments, we train for 10 epochs as these sources are prone to be overfitted. We use L1-loss that is computed on the waveform domain, optimized with Adam optimizer. Learning rate is set to 0.001 multiplied by a factor of 0.9 every 1.5 epochs. All our experiments are performed on 2 RTX3090 GPUs using fp32 precision.

5 Results and analysis

In this section, we first present experimental results on the MUSDB18 dataset, comparing the proposed method with state-of-the-art baselines. We then conduct an ablation study to verify the effectiveness of the components in the proposed method. Finally, we provide the MOS evaluation for the proposed method and several baselines.

5.1 Comparison with previous methods

To verify the advancement of our proposed model, we compare the proposed TFSWA-ResUNet with several existing state-of-the-art systems on MUSDB18 dataset in terms

of SDR. The comparison methods include D3Net [39], ResUNetDecouple+ [17], HT Demucs [33], CDE-HTCN [12], CWS-PResUNet [21], BS-RoFormer [24] and Mel-RoFormer [42]. The details of the baseline models are given as follows:

D3Net: A variant of DenseNet that incorporates multi-dilated convolutions. In this architecture, a novel multi-dilated convolution with different dilation factors within a single layer is employed, effectively avoiding the aliasing problem that can occur with a naive incorporation of dilation in the DenseNet architecture.

ResUNetDecouple+: A 143-layer residual UNet that decouples the estimation of magnitudes and phases. It enhances flexibility in magnitude estimation by combining bounded magnitude masks with direct prediction techniques.

HT Demucs: It employs a dual UNet structure, with temporal and spectral branches each having their respective skip connections. This allows the model to use whichever representation is most convenient for different parts of the signal and to freely share information between the two representations.

CDE-HTCN: It consists of a cross-domain encoder (CDE) that leverages features from both the time domain and the spectrogram domain, as well as a powerful hierarchic temporal convolutional network (HTCN) for the separation of multiple music sources.

CWS-PResUNet: It is a channel-wise subband phase-aware ResUNet, in which mixture signals are decomposed into subbands and then transformed into spectrograms via STFT. It uses an unbound complex ideal ratio mask to estimate each source, as described in [17]. The utilization of the channel-wise subband feature may limit unnecessary global weight sharing on the spectrogram and reduce computational resource consumption.

BS-RoFormer: It is a frequency-domain approach that integrates a band-split module with a hierarchical transformer architecture. The band-split module initially projects the input complex spectrogram into subband-level representations. Subsequently, a stack of hierarchical transformers processes these representations to model both inner-band and inter-band sequences, which are then used for multi-band mask estimation.

Mel-RoFormer: It replaces BS-RoFormer's empirically defined, non-overlapping band-split module with a Mel-band projection module. This Mel-band scheme maps frequency bins into overlapped subbands according to the perceptually-motivated mel scale.

Table 1 provides SDRs and the number of parameters for both the proposed model and previous methods. In the 'Domain' column, 'T' and 'F' indicate that the method is

Table 1 Comparison of SDRs and parameter quantity between existing methods and the proposed model

Method	Domain	Vocals	Drums	Bass	Other	Params.
D3Net [39]	F	7.24	7.01	5.25	4.53	7.9M
ResUNetDecouple+ [17]	F	8.98	6.62	6.04	5.29	102.0M
HT Demucs [33]	T+F	7.93	7.94	8.48	5.72	41.4M
CDE-HTCN [12]	T+F	7.37	7.33	7.92	4.92	12.6M
CWS-PResUNet [21]	F	8.9	6.38	5.93	5.84	202.6M
BS-RoFormer [24]	F	11.36	**10.27**	**11.15**	7.08	93.4M
Mel-RoFormer [42]	F	**11.60**	9.34	-	**7.93**	94.8M
TFSWA-ResUNet (proposed)	F	9.16	6.68	6.15	5.56	44.3M

waveform based and spectrogram based, respectively, while 'T + F' indicates that the corresponding method is a hybrid waveform/spectrogram-domain model. It can be seen that HT Demucs and CDE-HTCN are hybrid domain methods, and the proposed model along with the other comparison methods are all spectrogram-based source separation methods. To ensure a fair and consistent comparison, all performance metrics reported in Table 1 for the listed methods, including HT Demucs, are based on evaluations conducted without the use of any additional external training data. The first row shows the performance of D3Net [39] which achieves slightly lower SDRs than other existing systems but with the fewest parameters. The ResUNetDecouple+ method [17] achieves a *vocals* SDR of 8.98 dB among the compared methods, with 102 M parameters. The third to the fifth rows show the results of HT Demucs [33], CDE-HTCN [12], and CWS-PResUNet [21]. Among the compared methods, HT Demucs achieves relatively high *drums* and *bass* SDRs of 7.94 dB and 8.48 dB, respectively.

As in the last row of Table 1, the proposed TFSWA-ResUNet achieves the relatively high *vocals* SDR of 9.16 dB, surpassing HT Demucs' SDR of 7.93 dB. Compared to the existing single domain UNet architecture models such as CWS-PResUNet and ResUNetDecouple+, the proposed TFSWA-ResUNet method also achieves better SDR performance in separating *bass* and *drums*. Furthermore, the parameter size of TFSWA-ResUNet is only 44.3M, which is 22.9% of CWS-PResUNet and 43.4% of ResUNetDecouple+. The SDR for *bass* of the proposed method is lower compared to CDE-HTCN which has fewer parameters, which is possibly due to the limited frequency resolution of the *bass* spectrum for frequency-domain methods.

Compared to our proposed model, BS-RoFormer and Mel-RoFormer achieved better separation performance. However, their architectural complexity introduces substantial computational overhead. Specifically, BS-RoFormer and Mel-RoFormer have parameter sizes of 93.4M and 94.8M, respectively, each of which exceeds twice the number of parameters of the proposed model. More critically, the training time for a single source using BS-RoFormer can extend to four weeks, utilizing 16 Nvidia A100-80GB GPUs. This significant computational demand inherently limits their generalizability and practical deployment, particularly in resource-constrained environments. In contrast, the proposed TFSWA-ResUNet requires only two days of training time per source on two Nvidia RTX 3090 GPUs. Overall, by integrating an attention mechanism into a traditional convolutional model in a highly efficient manner, the proposed TFSWA-ResUNet achieves a good balance between performance and computational cost, making it a feasible option for widespread adoption, especially in applications where computational resources are limited or rapid iteration cycles are critical.

5.2 Ablation study

We also conduct an ablation study to verify the effectiveness of each component in the TFSWA block of the proposed model. As shown in Table 2, the 'Baseline' in the first row refers to the version of our proposed model without bottleneck blocks. '+Conv' denotes using convolutional blocks as bottleneck blocks on the basis of the baseline. Here, the bottleneck consists of four convolutional blocks, and each convolutional block has the same structure as that in Fig 2. When using TSA and FSA separately as bottleneck blocks, as shown in the third and fourth rows, the SDR performance for

Table 2 SDR values of ablation study for elements in the TFSWA block

Method	Vocals	Drums	Bass	Other	All	FLOPs
Baseline	8.67	6.28	5.84	5.06	6.46	551G
+Conv	8.91	6.47	5.98	5.24	6.65	605G
+TSA	8.81	6.54	5.92	5.11	6.60	554G
+FSA	8.86	6.41	5.79	5.45	6.63	554G
+TSA+FSA	9.05	6.58	5.99	5.51	6.78	558G
+TFSWA (proposed)	**9.16**	**6.68**	**6.15**	**5.56**	**6.89**	564G

Table 3 Comparison of the performance on vocals using different configurations. The inference time is measured on two RTX 3090 GPUs with a 30-second input audio

# Head	# TFSWA	# Channel	SDR	SIR	SAR	Inference Time
1	1	384	8.79	14.86	8.12	1.08 s
1	4	384	9.01	14.92	8.25	1.29 s
4	**4**	384	**9.16**	**15.09**	**8.45**	1.25 s
4	4	192	8.56	14.77	7.89	**1.03 s**
4	16	384	9.03	14.98	8.28	1.45 s
16	4	384	9.08	15.01	8.36	1.21 s

each track and the overall SDR (the last column in the table) are both improved compared to the Baseline UNet.

We see that using Conv blocks as bottleneck blocks achieves better SDR than using TSA or FSA separately. However, the model size and computational complexity (FLOPs) have increased significantly. When using serial TSA and FSA as intermediate blocks, it achieves a higher SDR than using TSA or FSA or convolutional bottleneck blocks separately, indicating the effectiveness of combining TSA and FSA. The highest SDR performance is achieved when introducing the TFSWA module as a bottleneck block, which demonstrates the effectiveness of the Swin transformer and its combination with TSA and FSA in the proposed model.

To further demonstrate the effectiveness of the proposed model, we also experiment with different combinations of hyperparameters and study the impact on the SDR, SIR, SAR and inference time for *vocals* in Table 3. The hyperparameters include the number of heads in multi-head self-attention (# Head), the number of TFSWA blocks (# TFSWA), and the number of final output channels of the encoder (# Channel). We observe that using only one TFSWA block with single-headed self-attention results in an SDR of 8.79 dB, an SIR of 14.86 and an SAR of 8.12 dB. Increasing the number of TFSWA blocks to 4 improves the SDR by 9.01 dB. Further gains are achieved when using 4-head MSA, resulting in an SDR of 9.16 dB, an SIR of 15.09 and an SAR of 8.45 dB, respectively. When we halve the number of output channels of the encoder, the inference time is decreased but the SDR is also reduced by 0.6 dB. When we increase the number of TFSWA blocks or the number of heads in MSA to 16, both SDR, SIR and SAR performance decline. Therefore, setting the number of TFSWA block and that of heads in MSA to 4 is sufficient to achieve the best SDR and SAR performance.

Table 4 Mean opinion score results when asking to rate the quality and absence of artifacts in the generated samples, from 1 to 5 (5 being the best grade)

Method	Vocals	Drums	Bass	Other	All
Ground Truth	4.79	4.67	4.63	4.51	4.65
ResUNetDecouple+ [17]	3.65	2.68	2.47	2.94	2.93
CDE-HTCN [12]	2.96	**3.03**	**3.05**	2.36	2.85
TFSWA-ResUNet(proposed)	**3.97**	2.79	2.55	**3.41**	**3.18**

5.3 Subjective evaluation

We also conduct mean opinion score evaluation following the approach in [8]. We asked 36 participants with professional backgrounds in the music fields to evaluate audio segments based on two aspects: audio quality and the presence of artifacts. Before the evaluation, the normal hearing of all participants was confirmed. We carefully prepared 48 samples music segments. These segments were selected from the separated outputs of ResUNetDecouple+, CDE-HTCN, our proposed TFSWA-ResUNet model, and the original ground truth from the MUSDB18 test set, containing *vocals*, *drums*, *bass* and *other* tracks. The original mixture was also provided to the participants as a reference signal in the evaluation. Following the recommendation in [35], we set the duration of each music segment to 10 s. This length was chosen to mitigate listener fatigue and promote the robustness and stability of their responses. Participants completed the evaluation in a quiet environment with minimal background noise, and were specifically instructed to use Sony WH-1000XM4 headphones, to ensure a controlled auditory experience.

In Table 4, we observe that compared to the best previous method ResUNetDecouple+, our proposed TFSWA-ResUNet has better MOS on all tracks, indicating that the separated sources by the proposed method also provide better auditory effect. Besides, the CDE-HTCN model achieves better mean opinion score in *drum* and *bass* tracks than the proposed method, which is consistent with the SDR performance results shown in previous experiments.

6 Conclusion

In this paper, we proposed a novel TFSWA-ResUNet for music source separation, an effective way of modeling music spectrogram features. We proposed to construct a new bottleneck block for the ResUNet architecture, in which the global time sequence attention, global frequency sequence attention and local shifted window attention are combined to capture the correlations of music spectrogram features. Experimental results show that the combination of global and local attentions yields a lightweight model but relatively good performance, highlighting the proposed method's potential as a feasible and widely adoptable solution, particularly in resource-constrained scenarios. For future work, we will explore the combination of waveform and spectrogram-domain methods and further enhance the SDR for *bass* and *drum* tracks.

Acknowledgements
We express our sincere appreciation to the anonymous reviewers, whose constructive feedback has significantly improved the quality of this manuscript.

Author Contributions
All authors contributed to the study conception and design. Z.Y. is responsible for running the experiments and writing the original manuscript. Y.S. revised the paper and performed results analysis. H.Y. and Y.Z. made many useful comments and suggestions on the manuscript. X. W. is responsible for giving work directions and revising the manuscript.

Funding
This work is partially supported by the Xi'an Science and Technology Plan Project, China Grant (No. 24GXFW0009), the Key Laboratory of the Ministry of Culture and Tourism (No.2024-09, No.2023-02),
the Fundamental Research Funds for the Central Universities, China Grant (No.GK202205035, No. GK202101004, No. GK202407007), the National Natural Science Foundation of China (No. 62377034), the Shaanxi Key Science and Technology Innovation Team Project (No. 2022TD-26).

Data Availability
The data are publicly available. MUSDB18 dataset is available at https://sigsep.github.io/datasets/musdb.html#musdb18-compressed-stems

Declarations

Conflict of interest
The authors declare no conflict of interest.

Received: 20 January 2025 Accepted: 25 August 2025
Published online: 02 September 2025

References

1. J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization. ArXiv preprint arXiv:1607.06450 (2016)
2. E. Cano, D. FitzGerald, A. Liutkus et al., Musical source separation: An introduction. IEEE Signal Process. Mag. **36**(1), 31–40 (2019)
3. J. Chen, Y. Lu, Q. Yu, et al. Transunet: Transformers make strong encoders for medical image segmentation. ArXiv preprint arXiv:2102.04306 (2021)
4. W. Choi, S.L. Walsh, M. Kawasaki, Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation. In: Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020) (2020)
5. W. Choi, M. Kim, J. Chung, et al., Lasaft: Latent source attentive frequency transformation for conditioned source separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 171–175 (2021)
6. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
7. J.S. Downie, Music information retrieval. Ann. Rev. Inf. Sci. Technol. **37**(1), 295–340 (2003)
8. A. Défossez, Hybrid spectrogram and waveform source separation. In: Proceedings of the ISMIR 2021 Workshop on Music Source Separation, pp 11–18 (2021)
9. A. Défossez, N. Usunier, L. Bottou, et al., Music source separation in the waveform domain. ArXiv preprint arXiv:1911.13254 (2019)
10. D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus). ArXiv preprint arXiv:1606.08415 (2016)
11. R. Hennequin, A. Khlif, F. Voituret et al., Spleeter: A fast and efficient music source separation tool with pre-trained models. J. Open Sour. Softwa. **5**(50), 2154 (2020)
12. Y. Hu, Y. Chen, W. Yang et al., Hierarchic temporal convolutional network with cross-domain encoder for music source separation. IEEE Signal Process. Lett. **29**, 1517–1521 (2022)
13. A. Jansson, E.J. Humphrey, N. Montecchio, et al., Singing voice separation with deep u-net convolutional networks. In: International Society for Music Information Retrieval Conference, pp 745–751 (2017)
14. J. Kim, H.G. Kang, Contrastive learning based deep latent masking for music source separation. Interspeech **2023**, 3709–3713 (2023)
15. M. Kim, W. Choi, J. Chung, et al., Kuielab-mdx-net: A two-stream neural network for music demixing. In: Proceedings of the MDX Workshop at ISMIR, Online (2021)
16. M. Kim, J.H. Lee, S. Jung, Sound demixing challenge 2023 music demixing track technical report: Tfc-tdf-unet v3. ArXiv preprint arXiv:2306.09382 (2023)
17. Q. Kong, Y. Cao, H. Liu, et al., Decoupling magnitude and phase estimation with deep resunet for music source separation. In: Proceedings of the 22nd International Conference on Music Information Retrieval (ISMIR), pp 342–349 (2021)
18. T. Li, J. Chen, H. Hou, et al., Sams-net: A sliced attention-based neural network for music source separation. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp 1–5 (2021)
19. L. Lin, Q. Kong, J. Jiang, et al., A unified model for zero-shot music source separation, transcription and synthesis. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR) (2021)
20. H. Liu, L. Xie, J. Wu et al., Channel-wise subband input for better voice and accompaniment separation on high resolution music. Interspeech **2020**, 1241–1245 (2020)
21. H. Liu, Q. Kong, J. Liu, Cws-presunet: Music source separation with channel-wise subband phase-aware resunet. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR) (2021a)

22. Z. Liu, Y. Lin, Y. Cao, et al., Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 9992–10002 (2021b)
23. F. Lluís, J. Pons, X. Serra, End-to-end music source separation: Is it possible in the waveform domain? Interspeech **2019**, 4619–4623 (2019)
24. W.T. Lu, J.C. Wang, Q. Kong, et al., Music source separation with band-split rope transformer. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 481–485 (2024)
25. Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans.on Audio, Speech, Language Process. **27**(8), 1256–1266 (2019)
26. Y. Luo, J. Yu, Music source separation with band-split RNN. IEEE/ACM Trans. Audio, Speech, Language Process. **31**, 1893–1901 (2023)
27. T. Nakano, K. Yoshii, Y. Wu, et al., Joint singing pitch estimation and voice separation based on a neural harmonic structure renderer. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp 160–164 (2019)
28. A.A. Nugraha, A. Liutkus, E. Vincent, Multichannel music separation with deep neural networks. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp 1748–1752 (2016)
29. A. van den Oord, S. Dieleman, H. Zen, et al., Wavenet: A generative model for raw audio. In: 9th ISCA Speech Synthesis Workshop (SSW) (2016)
30. J. Pons, J. Serrá, S. Pascual, et al., Upsampling layers for music source separation. In: 2023 31st European Signal Processing Conference (EUSIPCO), pp 311–315 (2023)
31. Z. Rafii, A. Liutkus, F.R. Stöter, et al., MUSDB18: A corpus for music separation. https://doi.org/10.5281/zenodo.1117372 (2017)
32. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. Med Image Comput Computer-Assisted Intervention - MICCAI **2015**, 234–241 (2015)
33. S. Rouard, F. Massa, A. Défossez, Hybrid transformers for music source separation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1–5 (2023)
34. R. Sawata, S. Uhlich, S. Takahashi, et al., All for one and one for all: Improving music separation by bridging networks. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 51–55 (2021)
35. B. Series, Method for the subjective assessment of intermediate quality level of audio systems. In: Proceedings of the International Telecommunication Union Radiocommunication Assembly, iTU-R BS.1534-3 (2015)
36. D. Stoller, S. Ewert, S. Dixon, Wave-u-net: A multi-scale neural network for end-to-end audio source separation. In: Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), pp 334–340 (2018)
37. F.R. Stöter, S. Uhlich, A. Liutkus et al., Open-Unmix: A reference implementation for music source separation. J. Open Sour. Softw. **4**(41), 1667 (2019)
38. F.R. Stöter, A. Liutkus, N. Ito, The 2018 signal separation evaluation campaign. In: International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pp 293–305 (2018)
39. N. Takahashi, Y. Mitsufuji, Densely connected multidilated convolutional networks for dense prediction tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 993–1002 (2021)
40. A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp 6000—-6010 (2017)
41. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
42. J.C. Wang, W.T. Lu, M. Won, Mel-band roformer for music source separation. In: Extended Abstracts for the Late Breaking Demo Session of the 24th Int. Society for Music Information Retrieval Conf (2023)
43. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(3), 483–492 (2016)
44. Q. Wu, H. Deng, K. Hu et al., Music source separation via hybrid waveform and spectrogram based generative adversarial network. Multimedia Tools and Applications (2024). https://doi.org/10.1007/s11042-024-20038-9
45. J.R. Zapata, M.E.P. Davies, E. Gomez, Multi-feature beat tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(4), 816–825 (2014)

Publisher's Note