Improved Wave-U-Net Blind Source Separation Using Content-Aware Filtering and Dynamic Downsampling

Sainan Tang, *Fei Long,* Shengbo Hu *a*nd Quan Liu

***Abstract*—This paper proposes Wave-U-Net+CAFDDS, an enhanced variant of Wave-U-Net that improves the performance of Blind Source Separation. To mitigate aliasing and feature loss in Wave-U-Net's decimation layers, we introduce two key enhancements. First, we replace all decimation layers with dynamic downsampling (DDS) layers. DDS adaptively selects sampling positions based on input features, thereby enhancing the model's ability to retain important features. Second, we insert a content-aware filter (CAF) before each downsampling stage. The CAF dynamically modulates its parameters according to feature context, reducing aliasing artifacts and boosting separation quality. We assess our model on the MUSDB18HQ dataset. Experimental results demonstrate that Wave-U-Net+CAFDDS, integrating both CAF and DDS, significantly outperforms the original Wave-U-Net.**

***Index Terms*—Blind Source Separation, decimation, content-aware filter, dynamic downsampling.**

I. INTRODUCTION

IN the field of signal processing, many techniques can be employed to extract the target signal according to the specific situation of the signal, such as noise suppression [1], [2], spectral subtraction, and filtering methods. These methods are typically focused on extracting a single source signal. For separating multiple sound sources, Blind Source Separation (BSS) techniques are commonly applied. BSS recovers individual signals from a mixture without prior information about their composition. It is widely applied in image processing, audio analysis, and medical signal processing.

Because source signals are often highly correlated and mixtures usually include background noise and interference [3], traditional BSS models often depend on strict mathematical assumptions to separate the signals. Independent Component Analysis (ICA) [4] assumes that signals are non-Gaussian and mutually independent. It separates sources by maximizing the non-Gaussianity of estimated components. Independent Vector Analysis (IVA) [5] extends ICA for multi-channel signals, but it still struggles with complex nonlinear mixtures. Nonnegative Matrix Factorization (NMF) [6] requires input signals to be nonnegative. It factorizes a nonnegative matrix into the product of two or more nonnegative matrices. Maximum Likelihood Estimation (MLE) [7] simplifies the BSS problem by modeling the probability distribution of the source signals. It estimates the parameters that maximize the likelihood of the observed mixtures under this distribution, allowing reconstruction of the source signals. The strict signal requirements of traditional BSS methods limit their applicability in practical applications.

Unlike traditional BSS methods that depend on numerous mathematical assumptions, deep learning performs source separation automatically by learning patterns in mixed signals through neural networks. Applications of deep learning in BSS are generally categorized into frequency-domain and time-domain approaches. In [8]–[10], mixed signals are split into amplitude and phase components, and then separation is done using only the amplitude. This approach often ignores phase information, which impacts separation accuracy. To address these limitations, recent studies have increasingly focused on time-domain separation methods rather than frequency-domain approaches [11], [12]. Wave-U-Net [12] is a typical end-to-end model known for its strong separation performance. It processes raw waveforms directly, avoiding the phase loss typically found in frequency-domain methods.

To improve Wave-U-Net's separation capability, researchers have proposed various improvements. MHE0 [13] incorporates minimum hyperspherical energy regularization during training, which improves the source separation performance. RA-Wave-U-Net [14] introduces two architectural changes: replacing convolutional layers with residual units in the encoder and decoder, and integrating attention gating into the skip connections. These modifications help bridge the semantic gap and enhance the model's separation performance. However, Wave-U-Net still has limitations in separating mixed signals. Traditional models reduce temporal resolution through successive decimation, which can lead to the loss of important signal details. According to the Nyquist–Shannon sampling theorem, decimation can cause high-frequency components to be misinterpreted as low-frequency ones, resulting in time-domain aliasing. This aliasing results in signal distortion and reduced model performance [15]. To address these issues, this article proposes two enhanced modules, dynamic downsampling (DDS) and content-aware filtering (CAF), to improve separation performance in audio source separation tasks.

Manuscript received December 23, 2024; revised July 27, 2025.

Sainan Tang is a postgraduate student at the School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China (e-mail: 232200232058@gznu.edu.cn).

Fei Long is a professor at the College of Artificial Intelligence and Electrical Engineering, Guizhou Institute of Technology, Guiyang 550003, China (corresponding author to provide e-mail: feilong@git.edu.cn).

Shengbo Hu is a professor at the School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China (e-mail: hsb@gznu.edu.cn).

Quan Liu is a postgraduate student at the School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China (e-mail: 242200232077@gznu.edu.cn).

II. MODEL FRAMEWORK

A. Wave-U-Net

The traditional Wave-U-Net model employs the classic U-Net architecture to perform end-to-end audio signal separation directly in the time domain, without relying on spectral representations. The primary objective of the model is to separate a mixed time-domain waveform $W \in [-1,1]^{C\times L_{in}}$ into N independent source time-domain waveforms $S_n \in [-1,1]^{C\times T_{out}}$ where $n \in \{1,\cdots,N\}$. Here, C denotes the number of audio channels, while L_{in} and L_{out} represent the input and output waveform lengths, respectively [12]. As shown in Fig. 1, the traditional Wave-U-Net architecture mainly consists of an encoder, a decoder, and skip connections. The encoder contains multiple multi-scale downsampling blocks and is responsible for extracting high-level features from the input audio. At each layer, the temporal resolution is reduced by a factor of two relative to the preceding layer. In the decoder, the model upsamples the extracted features step by step using multi-scale upsampling blocks. This process gradually restores the high-resolution signals. The output layer then generates the separated source signals. Skip connections combine features from the upsampled and downsampled at the same scale and send them to the corresponding upsampling block. This skip connection mechanism reduces information loss during downsampling and improves the accuracy of source separation.

B. Improved Model Architecture of Wave-U-Net

The traditional Wave-U-Net employs decimation in its downsampling block to reduce feature resolution. Although decimation is computationally efficient for reducing resolution, it discards crucial audio information, thereby degrading the quality of the separated signals. To prevent decimation-induced information loss from impairing source separation and reconstruction, we introduce a dynamic sampling module. This module replaces the decimation operation and utilizes learnable attention weights to retain the key features of the audio signal adaptively. Continuous downsampling may treat high-frequency components as low-frequency ones, leading to time-domain aliasing and distortion in the separated signals. To address this issue, we insert a one-dimensional content-aware filter before each downsampling step. This filter suppresses aliasing and enhances the restoration of the original audio source.

The Wave-U-Net model takes a mixed time-domain waveform as input and produces multiple source waveforms. As shown in Fig. 2, the mixed time-domain audio is processed by r consecutive downsampling blocks to extract high-level features. Each downsampling operation halves the temporal resolution of its input features. The decoder then reconstructs the audio signal using r consecutive upsampling blocks. Each upsampling operation doubles the temporal resolution of the input features. Skip connections fuse features from corresponding downsampling and upsampling blocks at the same scale, providing detailed information for reconstruction. Each downsampling block comprises a 1D convolutional layer, a content-aware filter, and a dynamic downsampling operation. The convolutional layer includes a 1D convolution and a LeakyReLU activation. The upsampling block comprises a linear-interpolation upsampling layer followed by a convolution layer. The Concat operation at the downsampling block r merges its feature with decoder features at the same scale. The model's final layer produces the separated audio outputs. Decoder outputs pass through a 1D convolution and Tanh activation to predict the first $N-1$ sources. The N-th source is computed by subtracting the sum of the first $N-1$ predictions from the mixture.

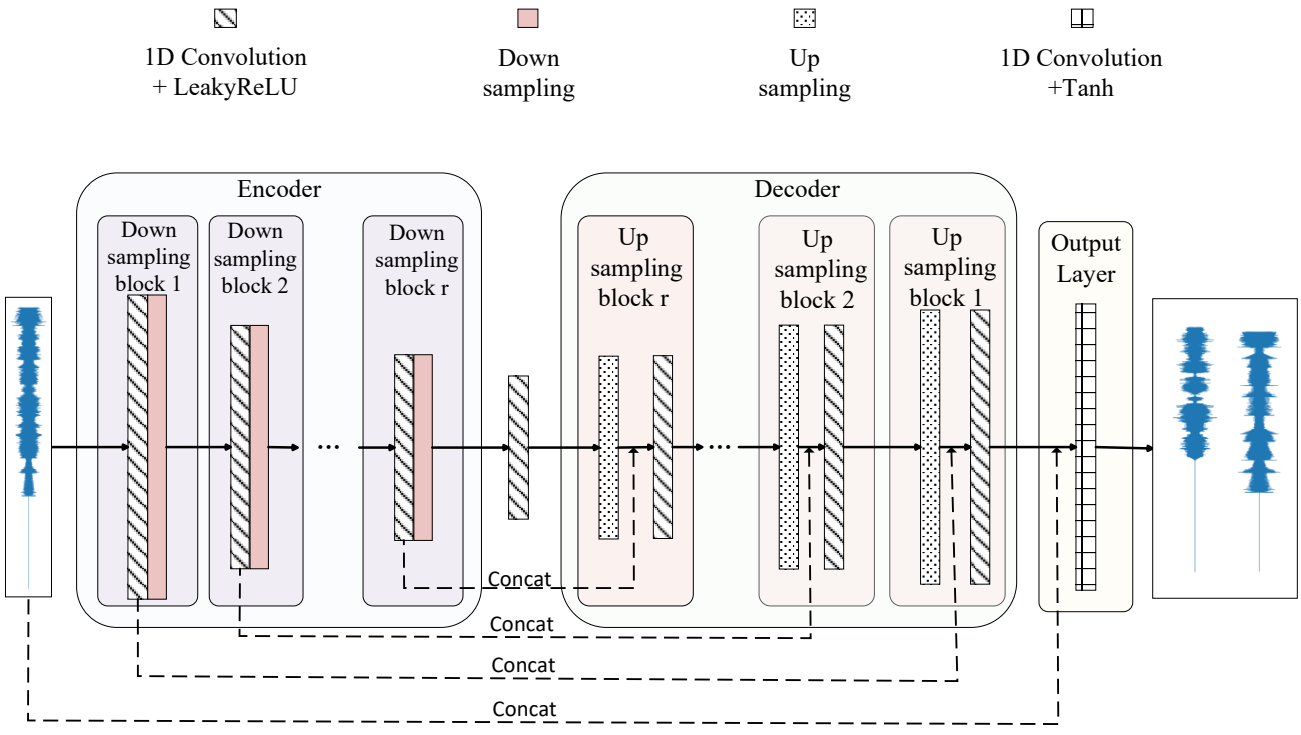


Fig. 1: Wave-U-Net model structure

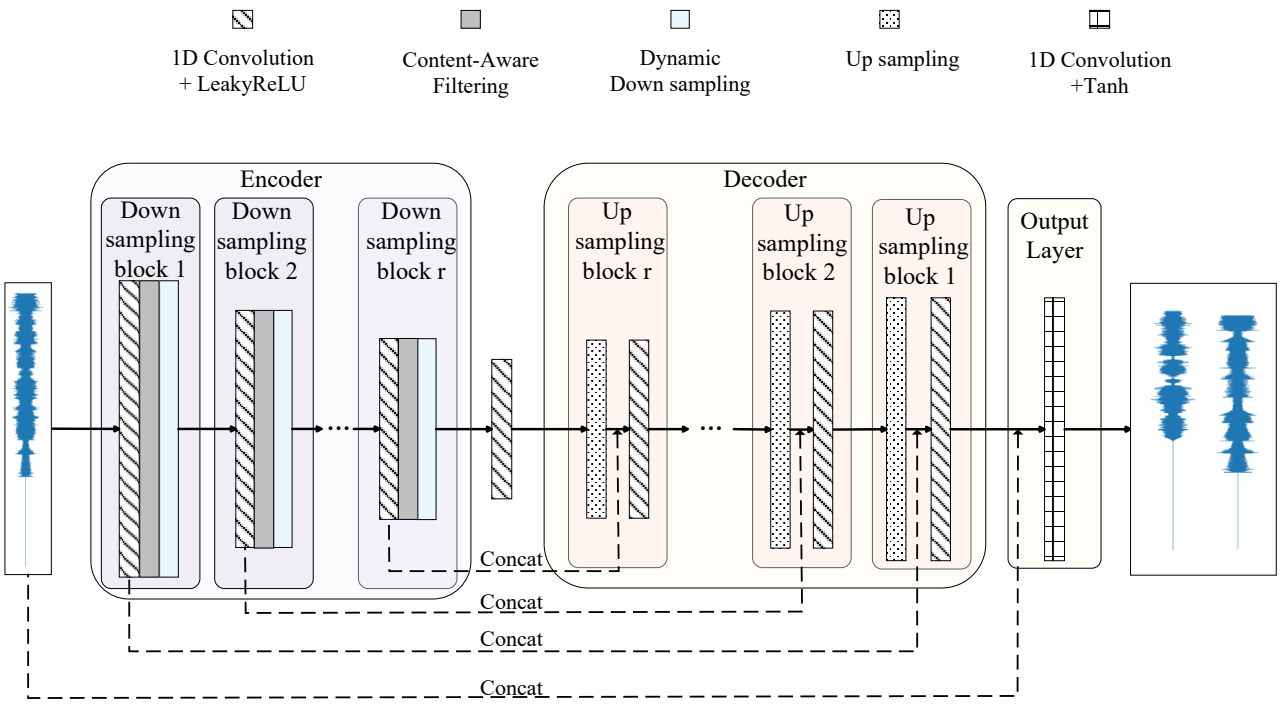


Fig. 2: Improved Wave-U-Net model structure

Table I presents the complete model architecture. $Conv1D(C,k)$ denotes a 1D convolutional layer with C input channels and a kernel size of k. $Concat(X)$ means concatenating the current high-level feature with the feature X. The parameter g denotes the number of feature groups, s is the downsampling factor, and r represents the number of downsampling and upsampling layers.

TABLE I: Details of the improved Wave-U-Net model structure

Block	Operation
Downsampling block r, r=1,…,R	Conv1D(C_d,k_d) Content-aware filtering(g) Dynamic downsampling(g,s)
Bridge layer	Conv1D(C_b,k_b)
Upsampling block r, r=R,…,1	Upsampling Concat(Dowsampling block r) Conv1D(C_u,k_u)
Output layer	Concat(input) Conv1D(C_i,k_i)

III. Methodology

A. Dynamic Downsampling

Traditional downsampling uniformly discards data at fixed intervals, potentially losing important information. In contrast, dynamic downsampling adaptively chooses sampling positions based on the input features. It focuses on task-relevant content and keeps key information. Inspired by the method in [16], this work applies dynamic sampling—initially developed for image processing—to one-dimensional time-domain audio waveforms.

Given an input feature $X \in R^{C\times L}$, we construct a standard sampling grid $G \in R^{g\times(L/s)}$, where C is the number of channels, L is the feature length, and s is the downsampling factor. After downsampling, the feature length becomes L/s. To reduce the computational cost of offset estimation, we split X along the channel dimension into g groups. We apply a 1D convolution to produce initial offsets $O \in R^{g\times(L/s)}$. Next, we introduce a Squeeze-and-Excitation (SE) [17] module. It uses time-domain global average pooling to extract channel information. The pooled vector passes through two fully connected layers (FC) and a Sigmoid activation to produce attention weights u. The SE module is defined as follows:

$$u = \sigma\left(F_2 \cdot \delta\left(F_1 \cdot GAP\left(X\right)\right)\right) \tag{1}$$

Where, $F_1 \in R^{(C/r)\times C}$ and $F_2 \in R^{C\times(C/r)}$ are the fully connected layer weights, r is the channel compression ratio, $GAP\left(\cdot\right)$ denotes global average pooling, δ is the ReLU activation, and σ represents the Sigmoid function.

We apply the attention weights to the initial offsets. Each offset is adjusted through element-wise multiplication and scaled by 0.5 to limit large shifts. The refined offsets are defined as follows:

$$offset = \left(conv1D\left(X\right) \odot conv1D\left(u\right)\right) \times 0.5 \tag{2}$$

Here, $\odot$ denotes element-wise multiplication.

Then, add the offsets to a standard sampling grid G to obtain the sampling grid P. Fig. 3 illustrates the dynamic offset generation process. The generation process of the sampling set can be expressed as:

$$P = offset\left(X\right) + G \tag{3}$$

Finally, use a grid sampling function to downsample the input feature X at the coordinates specified by the sampling grid P, producing the downsampled output $X_1 \in R^{C\times(L/s)}$:

$$X_1 = GridSample\left(X, P\right) \tag{4}$$

The dynamic downsampling process is shown in Fig. 4.

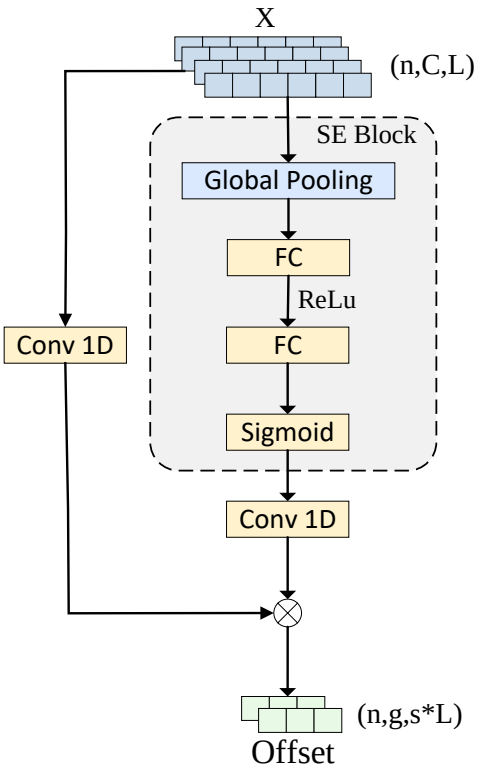


Fig. 3: Dynamic offset generation process

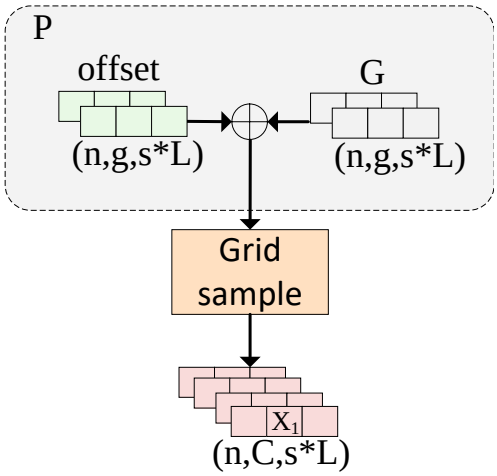


Fig. 4: Dynamic downsampling process

B. Content-Aware Filter

The conventional Wave-U-Net employs decimation to continuously downsample input features. This process can fold high-frequency components into lower frequencies, causing time-domain aliasing. The aliasing effect not only causes artifacts in the separated audio but also distorts the audio. Applying an anti-aliasing filter before downsampling can reduce this effect. In this work, we insert an anti-aliasing filter before each downsampling stage to prevent high-frequency folding, thereby reducing artifacts and distortion in the output audio. However, the energy distribution of audio signals varies greatly in different periods. Using a unified filter may not be able to capture the feature details of the input features throughout the entire time domain. To address this limitation, we design a one-dimensional content-aware filter based on the anti-aliasing framework in [18]. This filter adjusts its parameters according to local time segments of the audio. It helps the model adapt to signal changes and learn important features more effectively. The content-aware filter module is illustrated in Fig. 5.

Given an input feature $X \in R^{C\times L}$, where C is the number of channels and L is the sample length. To reduce computational complexity, we split X along the channel dimension into g groups. Each group then generates a filter applied to adjacent time-domain regions. Features within each group share a common filter. The filter weights $f_{i,g}^{p} \in R^{W}$ for the i-th group at time t as follows:

$$f_{t,i}^{p} = softmax\left(BN\left(conv1D\left(X\right)\right)\right) \tag{5}$$

Here, $softmax\left(\cdot\right)$ denotes the softmax function, $BN\left(\cdot\right)$ denotes batch normalization, and W is the local window size around position t. We apply the resulting filter weights to the corresponding group features. The filtering operation is defined as follows:

$$F_{t}^{i} = \sum_{p\in W} f_{t,i}^{p} \cdot X_{t+p}^{i} \tag{6}$$

Here, F_{t}^{i} denotes the output feature of the i-th group at time t; W is the local window centered at t; $f_{t,i}^{p}$ is the p-th filter weight at time t for group i; and X_{t+p}^{i} denotes the input feature of group i at position $t+p$

Since content-aware filters suppress high-frequency signals, they may lose important details. To retain useful information while minimizing aliasing, we employ a learnable residual fusion. The feature fusion mechanism is shown in Fig. 6. First, we modulate the filter's output via learnable parameters:

$$\tilde{F_{t}^{i}} = \sigma\left(\lambda\right) F_{t}^{i}, \sigma\left(\lambda\right) \in \left(0,1\right) \tag{7}$$

Next, we fuse the filtered and original features proportionally as follows:

$$X'^{i}_{t} = \alpha\left(\tilde{F_{t}^{i}}\right) + \left(1-\alpha\right) X_{t}^{i} \tag{8}$$

Here, $\sigma\left(\cdot\right)$ denotes the sigmoid activation function, and λ and α are learnable parameters.

IV. EXPERIMENTAL EVALUATION

A. Experimental Data and Environment Configuration

We employed the MUSDB18HQ dataset [19], which comprises 150 stereo tracks sampled at 44100 Hz, totaling approximately 10 hours of audio. Each track includes a mixture of its isolated sources (vocals, drums, bass, and accompaniment). This experiment focuses on the separation of vocal and accompaniment. We split the original 100-track training set into 75 tracks for training and 25 tracks for validation. We augmented the training data with the CCMixter dataset. CCMixter comprises 50 stereo tracks sampled at 44100 Hz that include mixes, vocals, and background sounds. We retained the original 50-track test set for evaluation. For each track, we randomly extracted 100 segments, 147443 samples in length, and downsampled them to 22050 Hz.

Table II lists the hardware and software configuration used in this experiment.

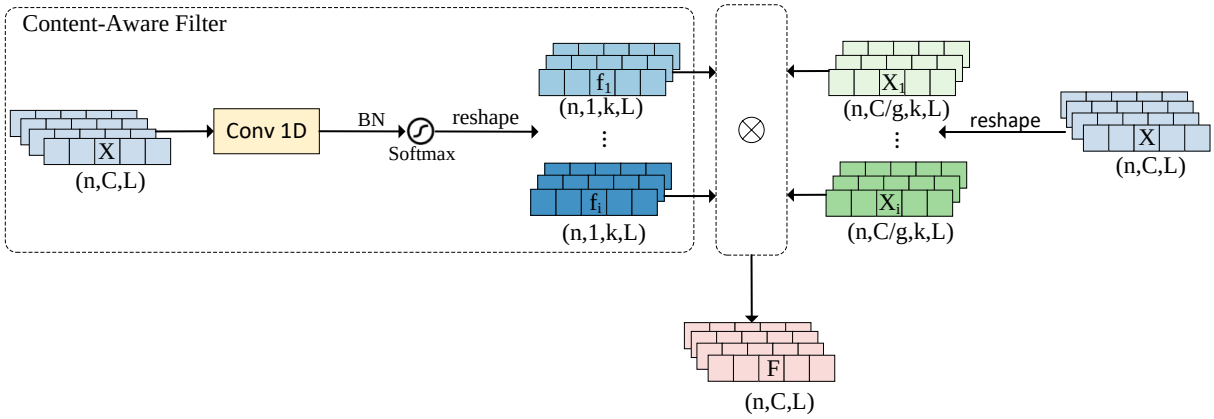


Fig. 5: Content-aware filter module structure

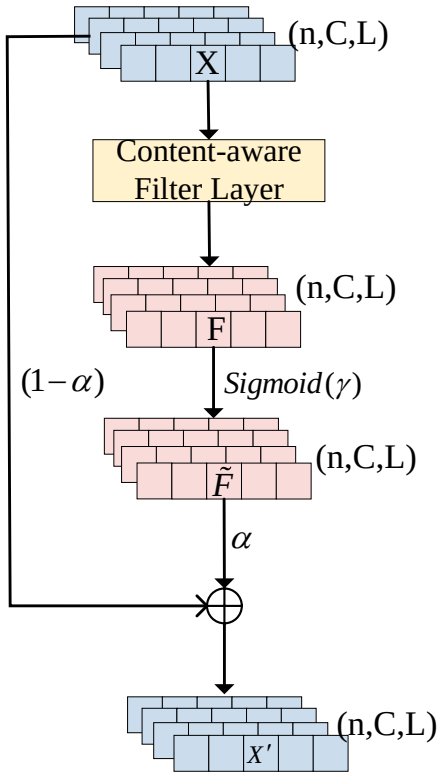


Fig. 6: Feature fusion mechanism

TABLE II: Details of experimental environment configuration

Configuration	Details
Programming Language	python 3.7
Framework	PyTorch 1.11
Development Environment	PyCharm
CUDA	CUDA 11.3
GPU	NVIDIA RTX A6000
CPU	Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz

B. Experimental Settings

1) Model Training and Setting

We trained the model using the Adam optimizer [20] with a learning rate of 0.0001 and a batch size of 16. The loss function was the Mean Absolute Error (MAE):

$$L_{MAE}(T_n, P_n) = \frac{1}{L_x}\sum_{t=1}^{L_x}|T_n - P_n| \tag{9}$$

Here, T_n denotes the ground-truth signal of the source, P_n is the predicted signal, and L is the total number of sampling points. Training stops if the validation loss does not improve for 20 consecutive epochs. We then select the model with the lowest validation loss as the final model. In this experiment, the network comprises 12 layers. Table III summarizes the input channels, output channels, and parameter configurations for each layer.

2) Evaluation Metrics

In Blind Source Separation, three metrics—signal distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR)—are commonly employed to assess model performance from distinct perspectives [21]. SDR quantifies the model's overall separation quality, SIR measures residual interference from other sources in the separated audio, and SAR assesses artifact levels introduced during separation. Their calculation formulas are as follows:

$$SDR = 10\log_{10}\frac{||S_{target}||^2}{||e_{interf}+e_{artif}||^2} \tag{10}$$

$$SIR = 10\log_{10}\frac{||S_{target}||^2}{||e_{interf}||^2} \tag{11}$$

$$SAR = 10\log_{10}\frac{||S_{target}+e_{interf}||^2}{||e_{artif}||^2} \tag{12}$$

In those formulas, S_{target} represents the target signal component extracted from the mixture signal, e_{interf} represents the interference error resulting from incomplete signal separation, and e_{artif} represents the artifacts or distortions introduced by the model.

In this experiment, we computed the mean and median of three metrics—SDR, SAR, and SIR—on the test set to evaluate model separation performance. The mean reflects the model's overall separation capability but is sensitive to outliers produced during source separation; relying solely on it may not represent typical performance. Therefore, we use the median as a complementary measure, reflecting performance in at least half of the separation segments and providing a more comprehensive evaluation.

TABLE III: The parameter setting of the improved Wave-U-Net model

Block	Operation	Input Channel	Oupt Channel
Dowsampling block r, r=1,…,12	Conv1D($k_d = 15$) Content-aware filtering($g = 2$) Dynamic downsampling($g = 2$, $s = 2$)	C_d=2,24,48,72,96,120, 144,168,192,216,240,264	C_o=24,48,72,96,120,144, 168,192,216,240,264,288
Bridge layer	Conv1D($k_b = 15$)	C_b=288	C_o=312
Upsampling block r, r=12,…,1	Upsampling Concat(Dowsampling block i) Conv1D($k_u = 5$)	C_u=600,552,504,456,408,360, 312,264,216,168,120,72	C_o=288,264,240,216,192, 168,144,120,96,72,48,24
Output layer	Concat(input) Conv1D($k_i = 5$)	C_i=24	C_o=2

C. Ablation Experiment

To assess how the added modules affect separation performance in a Blind Source Separation task, we compared the baseline Wave-U-Net model with three enhanced variants. The first variant integrates a content-aware filter (CAF), the second employs dynamic downsampling (DDS), and the third combines both content-aware filtering and dynamic downsampling (CAFDDS). The results of this comparison are presented in Tables IV and V.

As shown in Fig. 7, the first quartile (Q1) marks the best separation result among the lowest 25% of segments. A higher Q1 means better performance in low-quality segments. The third quartile (Q3) shows the worst separation result among the top 25%. A higher Q3 indicates better quality in high-performing segments. A shorter lower whisker indicates stable performance among the lowest-performing segments. A longer upper whisker indicates that the model can reach a higher potential in the best-performing segments.

1) Analysis of Vocal Separation Performance

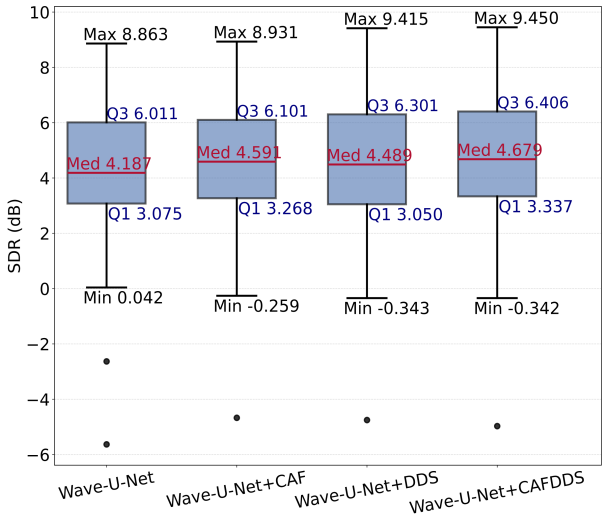
Table IV shows that for vocal separation, the baseline Wave-U-Net has the lowest mean SDR (2.450 dB) and mean SAR (4.015 dB). Its SAR median (4.486 dB) and lower whisker (–3.984 dB) in Fig. 7(e) are also the lowest among all models. These results suggest that the baseline model performs poorly in overall vocal separation and causes severe distortion in some segments. With the CAF module, Wave-U-Net+CAF exhibits the smallest interquartile range for SDR, SIR, and SAR, as seen in Fig. 7(a), (c), and (e). Its SAR mean (4.239 dB) and median (4.879 dB) both increase compared to the baseline, indicating fewer artifacts and a more concentrated distribution. However, the SIR median drops to 12.851 dB, suggesting weaker suppression of accompaniment interference.

In contrast, the dynamic downsampling variant (Wave-U-Net+DDS) improves the SIR, with a mean of 8.653 dB and a median of 14.528 dB. The SAR also increases slightly, reaching a mean of 4.241 dB and a median of 4.684 dB. Fig. 7(a), (c), and (e) show that the medians of SDR, SIR, and SAR, as well as Q3 and the upper whiskers, increase significantly. Q1 and the lower whiskers also decrease. These results show that, although DDS enhances accompaniment suppression and reduces vocal distortion, it also degrades performance on the lowest-quality segments, leading to greater variability in overall separation.

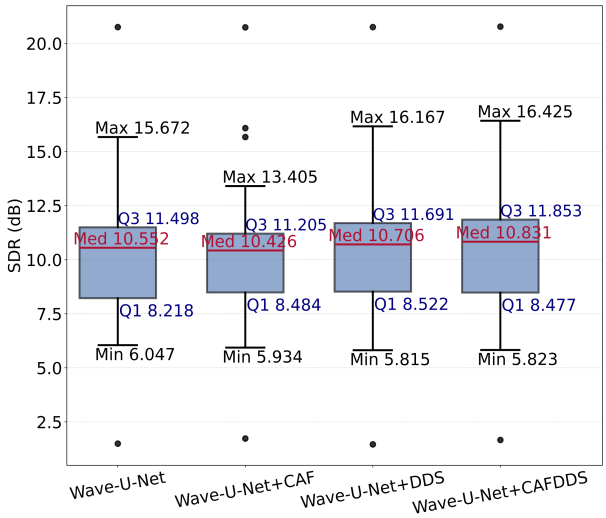
The combined model (Wave-U-Net+CAFDDS) achieves the highest mean and median SDR, increasing by 19.55% and 11.75%, respectively. It shows that CAF and DDS together keep better separation across most test audio. The model also achieves the highest mean SIR (8.858 dB), mean SAR (4.324 dB), and median SAR (5.230 dB), indicating improved interference suppression and fewer artifacts. Fig. 7(a), (c), and (e) show that SDR and SIR achieve the highest Q1 and Q3 values, while SAR exhibits the highest Q3 value. All boxes shift upward, and the SDR upper whisker reaches 9.450 dB, and the SIR upper whisker reaches 25.276 dB. These results show that the model achieves better separation than the baseline for most segments. Compared with using a single module, combining both modules reduces severe distortions and maintains high separation quality.

TABLE IV: Vocal separation performance across different models

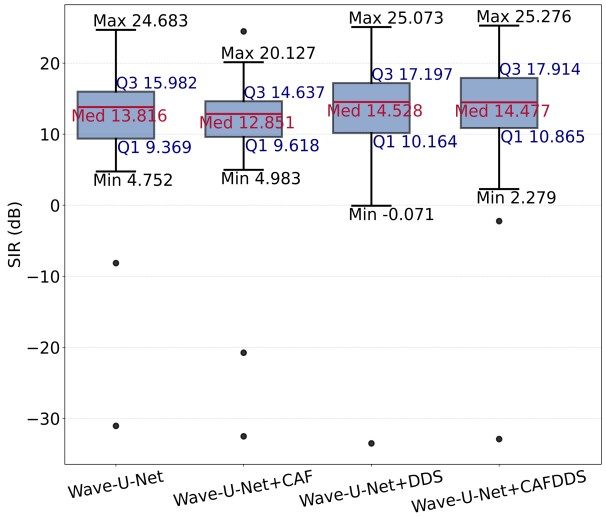
Model	CAF	DDS	Mean			Median		
			SDR	SIR	SAR	SDR	SIR	SAR
Wave-U-Net	×	×	2.450	7.851	4.015	4.187	13.816	4.486
Wave-U-Net+CAF	✓	×	2.363	6.688	4.239	4.591	12.851	4.879
Wave-U-Net+DDS	×	✓	2.841	8.653	4.241	4.489	14.528	4.684
Wave-U-Net+CAFDDS	✓	✓	2.929	8.858	4.324	4.679	14.477	5.230



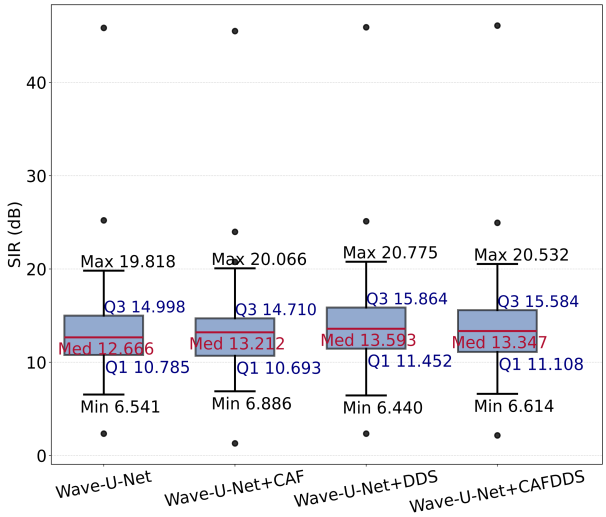
(a) Vocal separation SDR comparison



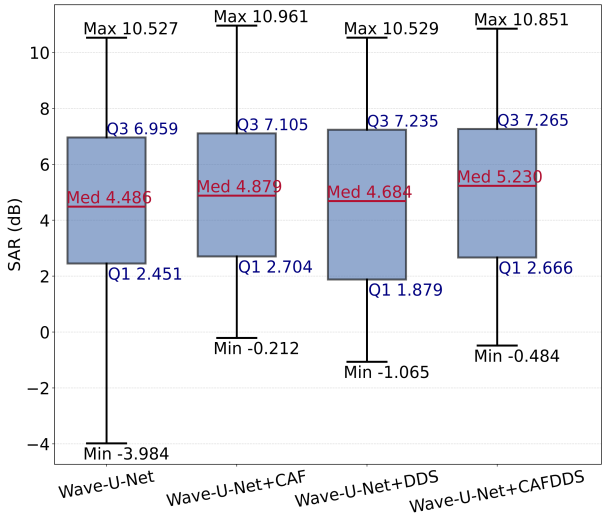
(b) Accompaniment separation SDR comparison



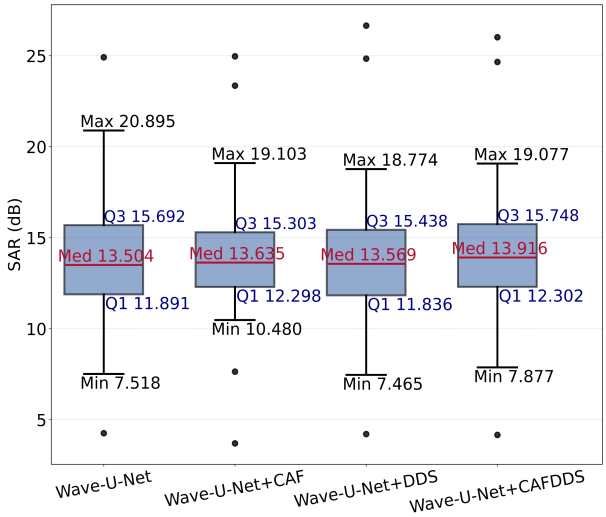
(c) Vocal separation SIR comparison



(d) Accompaniment separation SIR comparison



(e) Vocal separation SAR comparison



(f) Accompaniment separation SAR comparison

Fig. 7: Box plot comparison of separation metrics

TABLE V: Accompaniment separation performance across different models

Model	CAF	DDS	Mean			Median		
			SDR	SIR	SAR	SDR	SIR	SAR
Wave-U-Net	×	×	11.434	16.427	14.230	10.552	12.666	13.504
Wave-U-Net+CAF	✓	×	11.469	16.449	14.302	10.426	13.212	13.635
Wave-U-Net+DDS	×	✓	11.686	17.005	14.310	10.706	13.593	13.569
Wave-U-Net+CAFDDS	✓	✓	11.759	16.746	14.668	10.831	13.347	13.916

2) Analysis of Accompaniment Separation Performance

Table V shows that for accompaniment separation, Wave-U-Net+CAF improves the mean SDR by only 0.31%, while the median SDR drops by 1.19%. In Fig. 7(b), Q1 increases from 8.218 dB to 8.484 dB, but Q3 and the upper whisker decrease by 2.55% and 14.47%. These results suggest that CAF improves performance on low-quality samples but reduces it for middle and high-quality ones compared to the baseline. For SIR, Wave-U-Net+CAF shows a slight mean increase of 0.13% and a median increase of 4.31%. The lower whisker rises slightly, but Q1 and Q3 fall by 0.85% and 1.92% (Fig. 7(d)). It indicates that CAF improves the lowest interference cases but weakens performance for better ones. For SAR, both the mean and median increase slightly. Q1 rises by 3.42% and the lower whisker by 39.40%, while Q3 and the upper whisker drop by 2.48% and 8.60% (Fig. 7(f)). This result means that CAF enhances artifact suppression in poor samples but reduces it in high-quality ones.

The dynamic downsampling model (Wave-U-Net+DDS) outperforms the baseline in both SDR and SIR. Its SDR mean and median rise by 2.20% and 1.46%, respectively, with Q1, Q3, and the upper whisker all shifting upward (Fig. 7(b)), indicating improved separation for most samples. SIR mean and median rise by 3.52% and 7.32%, and both Q1 and Q3 increase (Fig. 7(d)), indicating more potent suppression of interference. SAR mean and median increase slightly, but Q3 and the upper whisker decrease (Fig. 7(f)), suggesting limited improvement in reducing artifacts.

The combined CAF and DDS variant (Wave-U-Net+CAFDDS) increases the SDR mean and median by 2.84% and 2.64%, respectively, compared to the baseline. Its Q1 exceeds the baseline, with Q3 and the upper whisker reaching maxima (11.853 dB and 16.425 dB), and only a slight drop in the lower whisker (Fig. 7(b)), indicating superior overall separation quality. For SIR, the mean and median rise by 1.94% and 5.38%, respectively; all quartiles and whiskers shift upward (Fig. 7(d)), indicating interference suppression comparable to DDS but with a more balanced distribution. For SAR, the mean and median rise by 3.08% and 3.05%; Q1 increases by 3.46% and the lower whisker by 4.78% (Fig. 7(f)), indicating reduced severe distortions in poor samples.

In summary, the CAF module enhances artifact suppression, improves the performance of low-quality segments, and increases model stability. The DDS module raises the model's upper performance limit and strengthens interference suppression. Combining both, Wave-U-Net+CAFDDS achieves the highest mean and median SDR and improves results across all segment qualities. It shows the best overall performance.

D. Performance Comparison of Blind Source Separation Model Algorithm

Table VI shows that for vocal separation, Wave-U-Net+CAFDDS achieves the highest mean SDR, SIR, and SAR. Its SIR median reaches 14.477 dB. However, its median SDR and SAR are slightly lower than those of Demucs [22]. This result indicates that, in the vocal separation task, Wave-U-Net+CAFDDS is effective in suppressing accompaniment compared to Demucs, improving vocal quality, but remains less effective in reducing artifacts. Table VII shows that for accompaniment separation, Wave-U-Net+CAFDDS obtains the highest mean and median SDR and SAR. It suggests better artifact reduction and clearer outputs. However, its SIR is lower than that of Demucs, showing weaker suppression of vocal interference. In summary, Wave-U-Net+CAFDDS

TABLE VI: Comparison of vocal separation performance across various blind source separation models

Model	Mean			Median		
	SDR	SIR	SAR	SDR	SIR	SAR
Wave-U-Net	2.450	7.851	4.015	4.187	13.816	4.486
Demucs	2.731	8.189	4.055	4.722	14.286	5.353
MHE0	2.636	7.714	4.137	4.480	13.436	5.009
AR-Wave-U-Net	2.467	7.402	3.910	4.666	13.563	5.383
Wave-U-Net+CAFDDS	2.929	8.858	4.324	4.679	14.477	5.230

TABLE VII: Comparison of accompaniment separation performance across various blind source separation models

Model	Mean			Median		
	SDR	SIR	SAR	SDR	SIR	SAR
Wave-U-Net	11.434	16.427	14.230	10.552	12.666	13.504
Demucs	11.392	18.403	13.152	10.672	15.621	12.623
MHE0	11.625	16.738	14.267	10.647	13.267	13.466
AR-Wave-U-Net	11.250	17.907	13.178	10.592	14.930	12.597
Wave-U-Net+CAFDDS	11.759	16.746	14.668	10.831	13.347	13.916

achieves effective artifact suppression in both vocal and accompaniment separation tasks, resulting in the highest overall separation quality.

To quantify complexity, we compare each model's parameter count and the number of Multiplications and Accumulations (MACs), as summarized in Table VIII. Wave-U-Net+CAFDDS has a higher parameter count and MACs than Wave-U-Net and MHE0 but remains considerably leaner than Demucs and RA-Wave-U-Net. The added complexity stems from integrating new functional modules into Wave-U-Net. It trades increased computation for enhanced separation quality. Therefore, Wave-U-Net+CAFDDS is best suited for BBS tasks that require a high-quality output audio signal and have sufficient computational resources.

TABLE VIII: Comparison of computational complexity across various blind source separation models

Model	Params(M)	MACs(G)
Wave-U-Net	10.26	11.41
Demucs	265.68	39.22
MHE0	10.27	13.86
AR-Wave-U-Net	24	24.75
Wave-U-Net+CAFDDS	11.30	18.05

E. Visualization of Model Separation Results

To visualize the separation results, we randomly selected a mixture audio test segment and split it into vocal and accompaniment tracks. Fig. 8 presents the spectrogram of the mixture audio, and Fig. 9 presents the spectrograms of the separated vocal and accompaniment. The visualized results show that the proposed model improves noise suppression and produces higher-quality separated audio.

For accompaniment separation, Wave-U-Net fails to remove vocals above 2048 Hz entirely. Between 128 Hz and 512 Hz, the energy distribution is uneven, and the spectrogram contains many artifacts (Fig. 9(c)). In comparison, Wave-U-Net+CAFDDS removes abnormal energy above 2048 Hz and reduces low-frequency artifacts, resulting in cleaner output (Fig. 9(e)).

For vocal separation, Wave-U-Net shows low-frequency detail loss and intense background noise in silent regions (Fig. 9(d)). Wave-U-Net+CAFDDS+BN preserves more detail in low frequencies and suppresses noise during silence (Fig. 9(f)).

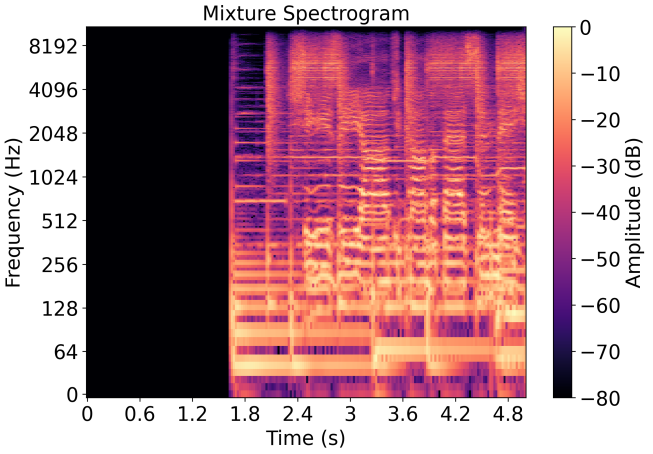
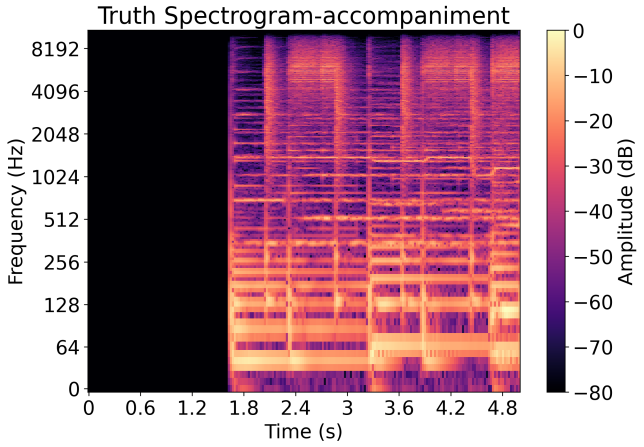


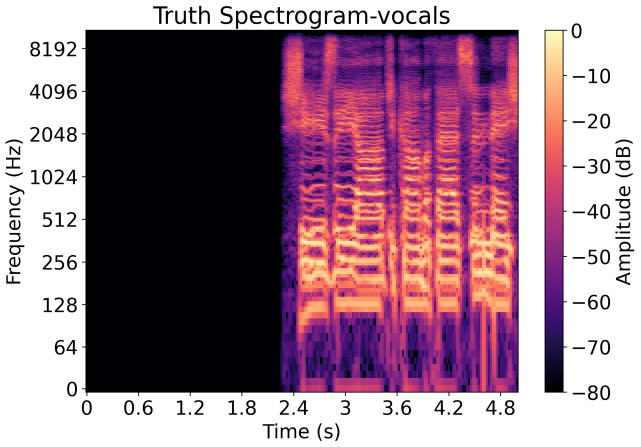
Fig. 8: Mixture audio spectrogram

V. CONCLUSION

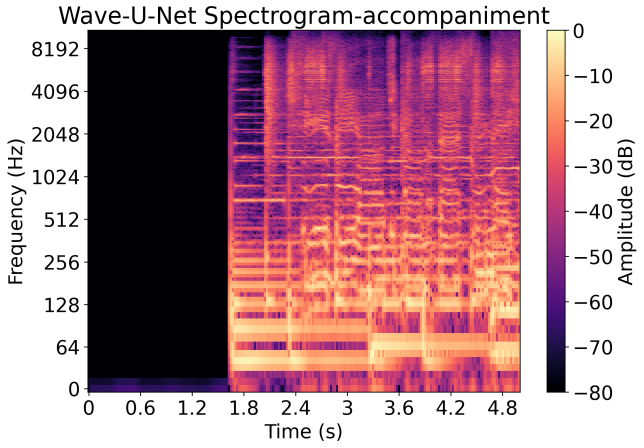
In this paper, we introduce Wave-U-Net+CAFDDS, an enhanced version of the traditional Wave-U-Net designed to improve the quality of separated source signals. To reduce aliasing caused by the decimation layer in the original downsampling process, we add a content-aware filter before downsampling. This filter adjusts adaptively to the input features, helping to keep important information and reduce the adverse effects of aliasing on performance. In the downsampling stage, the decimation layer is replaced with a dynamic downsampling layer, where dynamic sampling is combined with an SE block to constrain the sampling offset. The downsampling strategy varies according to the input characteristics, enabling the model to capture essential features better and retain more signal details. Experimental results on the MUSDBHQ18 dataset demonstrate that the proposed improvements enhance the model's performance. Compared to other blind source separation models, Wave-U-Net+CAFDDS delivers significant quality gains. It achieves the highest overall separation performance among all evaluated models. Future research will focus on two key directions: developing lightweight models and optimizing reconstruction quality. It includes reducing the number of downsampling layers and optimizing the upsampling layers. These efforts aim to improve further both the model's training speed and the quality of signal reconstruction.



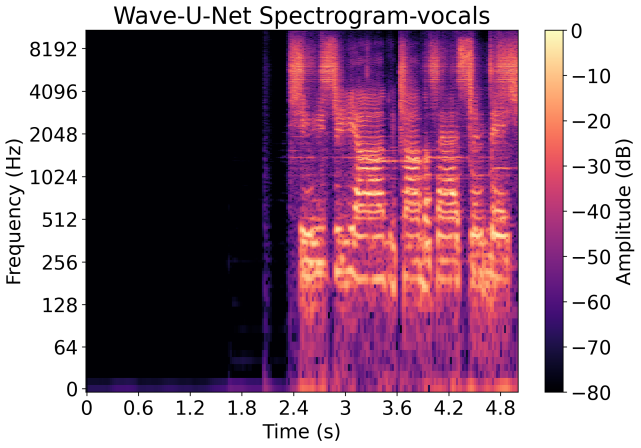
(a) True spectrum of the accompaniment



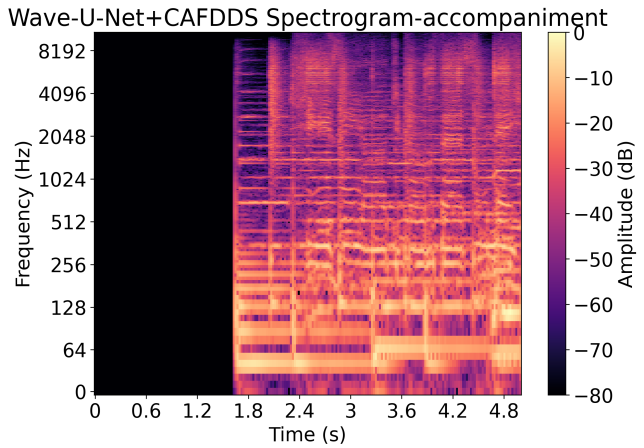
(b) True spectrum of the vocal



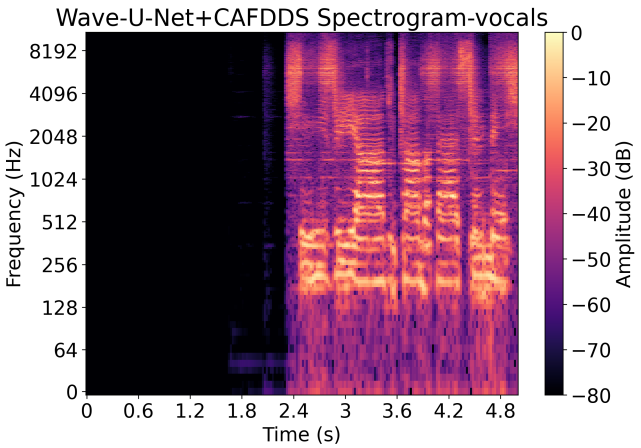
(c) Accompaniment spectrum separated by the traditional Wave-U-Net model



(d) Vocal spectrum separated by the traditional Wave-U-Net model



(e) Accompaniment spectrum separated by Wave-U-Net+CAFDDS model



(f) Vocal spectrum separated by Wave-U-Net+CAFDDS model

Fig. 9: Visualization of audio signal separation results

REFERENCES

[1] R. Guo, T. Jiang, Q. Wang, R. Liang, and C. Zou, "An Improved Low-Complexity Echo Suppression Algorithm Based on the Acoustic Coloration Effect," *IAENG International Journal of Computer Science*, vol. 49, no. 3, pp. 637–643, 2022.

[2] E. Verteletskaya and B. Simak, "Noise Reduction Based on Modified Spectral Subtraction Method," *IAENG International Journal of Computer Science*, vol. 38, no. 1, pp. 82–88, 2011.

[3] J. Kapoor, A. Pathak, M. Rai, and G. Mishra, "Speech Quality Enhancement through Noise Cancellation using an Adaptive Algorithm," *IAENG International Journal of Computer Science*, vol. 49, no. 3, pp. 653–665, 2022.

[4] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[5] R. Guo, Z. Luo, and M. Li, "A survey of optimization methods for independent vector analysis in audio source separation," *Sensors*, vol. 23, no. 1, p. 493, 2023.

[6] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[7] B. Pearlmutter and L. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ica," *Advances in Neural Information Processing Systems*, vol. 9, 1996.

[8] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," *2018 16th International workshop on acoustic signal enhancement (IWAENC)*, pp. 106–110, 2018.

[9] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate What You Describe: Language-Queried Audio Source Separation," *Interspeech 2022*, pp. 1801–1805, 2022.

[10] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[13] J. Perez-Lapillo, O. Galkin, and T. Weyde, "Improving singing voice separation with the wave-u-net using minimum hyperspherical energy," *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3272–3276, 2020.

[14] B. Wang and N. Chen, "An End-to-End Singing Voice Separation Model Based on Residual Attention U-Net," *Journal of East China University of Science and Technology (Natural Science Edition)*, vol. 47, no. 5, pp. 619–626, 2021.

[15] Y. Gong and C. Poellabauer, "Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models," *Interspeech 2018*, pp. 2698–2702, 2018.

[16] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to Upsample by Learning to Sample," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6004–6014, 2023.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2019.

[18] X. Zou, F. Xiao, Z. Yu, Y. Li, and Y. J. Lee, "Delving deeper into anti-aliasing in convnets," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 67–81, 2023.

[19] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[20] D. Kinga, J. B. Adam *et al.*, "A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, vol. 5, no. 6, 2015.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.