

USE: A Unified Model for Universal Sound Separation and Extraction

Hongyu Wang^{1,2*}, Chenda Li^{1,2*†}, Xin Zhou^{1,2}, Shuai Wang^{3,4}, Yanmin Qian^{1,2‡}

¹Auditory Cognition and Computational Acoustics Lab

MoE Key Lab of Artificial Intelligence, AI Institute

School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

²VUI Labs, China

³Nanjing University, Suzhou, China

⁴Shenzhen Loop Area Institute, Shenzhen, China

{hongyu.wang, lichenda1996, zxzx818, yanminqian}@sjtu.edu.cn

{shuawaiwang}@nju.edu.cn

Abstract

Sound separation (SS) and target sound extraction (TSE) are fundamental techniques for addressing complex acoustic scenarios. While existing SS methods struggle with determining the unknown number of sound sources, TSE approaches require precisely specified clues to achieve optimal performance. This paper proposes a unified framework that synergistically combines SS and TSE to overcome their individual limitations. Our architecture employs two complementary components: 1) An Encoder-Decoder Attractor (EDA) network that automatically infers both the source count and corresponding acoustic clues for SS, and 2) A multi-modal fusion network that precisely interprets diverse user-provided clues (acoustic, semantic, or visual) for TSE. Through joint training with cross-task consistency constraints, we establish a unified latent space that bridges both paradigms. During inference, the system adaptively operates in either fully autonomous SS mode or clue-driven TSE mode. Experiments demonstrate remarkable performance in both tasks, with notable improvements of 1.4 dB SDR improvement in SS compared to baseline and 86% TSE accuracy.

Demo — <https://hongyuhwang414.github.io/USE-demo/>

Introduction

In complex acoustic environments, Universal Sound Separation(USS) is capable of providing high-quality inputs for subsequent audio analysis tasks. USS has become a core task (Kavalerov et al. 2019; Tzinis et al. 2020, 2022; Liu et al. 2024b; Pons et al. 2024; Zhao et al. 2024a; Kong et al. 2023) and aims to isolate arbitrary types of sound sources, including speech, music, environmental sounds, and musical instruments, from complex audio mixtures without relying on predefined numbers or types of sound sources. This flexibility greatly expands the application scope of sound processing, making it indispensable in fields such as environmental monitoring and multimedia content analysis.

Despite recent advances in sound separation (SS) models, existing methods still face several limitations. For instance, many models require the number of sound sources

in a mixture to be predefined during inference (Luo and Mesgarani 2019; Zhao et al. 2024b; Subakan et al. 2021), which restricts their applicability in real-world scenarios where the number of sound sources is often uncertain. To address these issues, target sound extraction (TSE) utilizes some prior knowledge about the target sound from the mixture of an unknown number of sources (Liu et al. 2024a,b; Zhao et al. 2018; Liu et al. 2022; Kilgour et al. 2022; Zhang et al. 2024). The prior knowledge about the target sound, also known as *clue*. For example, the AudioSep model (Liu et al. 2024a) has shown remarkable performance by using the natural language description as the extraction clue; APT (Liu et al. 2024b) uses audio samples as the clue for the TSE task; PixelPlayer (Zhao et al. 2018) uses visual information to extract target sounds.

Although target sound extraction can theoretically solve the problem of the unknown number of sources in the mixture, it suffers from two drawbacks in real applications: 1) First, the clues that are assumed to exist in advance may be of low quality or not be found in actual applications, which will degrade the performance of the TSE or even extract the wrong target. 2) The second drawback is that some previous studies (Delcroix et al. 2020; Elminshawi et al. 2022) found that the TSE-based methods usually achieve less performance than the SS methods when the number of sources is determined.

In this paper, we propose Universal sound Separation and target sound Extraction (USE), a unified model, to solve the above-mentioned drawbacks of SS and TSE in complex acoustic scenarios. The USE comprises a separation backbone, an EDA module, and a multi-modal clue module. Without knowing the number of sound sources, the USE model can complete the SS task with a built-in EDA module. Following the previous work (Li et al. 2023), the multi-modal clue module is robust for missing or low-quality clues. It takes an uncertain number of clues of different modalities as input and generates a clue embedding in a unified space. When there is any available modality of clues, the USE model can accomplish the TSE task with the clue module. In the training stage, we constrain the EDA embeddings and clue embeddings into a unified space to align them in a

*These authors contributed equally.

†Corresponding authors

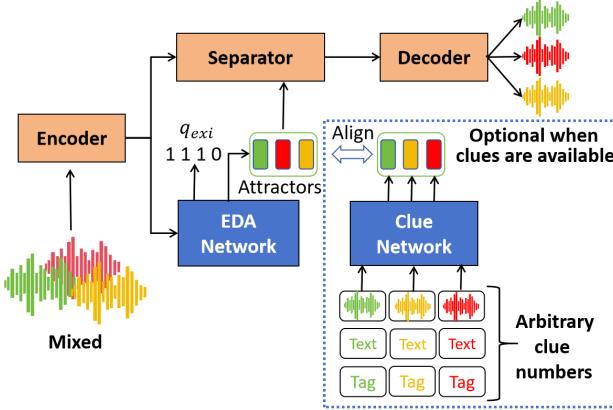


Figure 1: Architecture of the USE Framework. Without clues, the EDA Network implicitly learns to represent all independent sound sources in the mixture as attractors, which are then used for separation. With arbitrary clues, the Clue Network generates Clue embeddings that are sent into the separation network for extraction. During training, semantic alignment between Attractors and Clue embeddings bridges the gap between Sound Separation (SS) and Target Sound Extraction (TSE).

semantic space, which allows the USE model to perform SS or TSE tasks flexibly in the inference stage. We conducted the experiments on universal sound datasets with clues of multiple modalities, which shows the robustness of USE in complex acoustic scenarios.

The main novelty and contributions of this paper are as follows: 1) We propose USE, a unified framework for SS and TSE, which can separate or extract sounds from a mixture of an unknown number of sources; 2) We show that the multi-task USE brings better performance on both SS and TSE baselines. 3) We conduct the experiments on a universal sound separation dataset with low-quality clues of multiple modalities, which shows the robustness of USE in complex acoustic scenarios.

Related Work

There are some attempts to solve the problems of SS and TSE mentioned above. For example, OR-PIT (Takahashi et al. 2019) trains an independent binary classifier to determine whether speech should be further separated from the mixture; In (Li et al. 2023; Cheng et al. 2024), the impact of low-quality clues and clue missing are alleviated by utilizing the multi-modal clue; (Chetupalli and Habets 2022; Lee et al. 2024) proposed to use an Encoder-Decoder-Attractor to count the number of sources in the SS task. While these approaches have effectively improved performance in single-task SS and TSE, unifying these tasks remains a challenge. Although attempts like those in (Saijo et al. 2023; Chetupalli and Habets 2024) have used attractors to integrate SS and TSE, their methods are still relatively superficial. They fail to align attractors with clues in a semantic space, leading to a disjointed performance when handling both SS and TSE tasks, and they only focus on clean human

speech rather than universal sound.

The methods proposed in (Saijo et al. 2023) and (Chetupalli and Habets 2024) struggle to accurately associate the target sound sources with the given clues during the TSE task. In (Saijo et al. 2023), the TSE inference process involves weighting the attractors based on the weighted coefficients of the clues and attractors to extract the target audio. This method fails to effectively achieve semantic unification between attractors and clues. Additionally, it introduces extra computational overhead, particularly when dealing with a large number of sources in the mixture. Moreover, it relies solely on a single type of target source-related clue. When the quality of the clue is low or even absent, the TSE task cannot be effectively accomplished. In (Chetupalli and Habets 2024), although the TSE task incorporates classification learning of audio source in attractors, it still does not consider the substitutability between attractors and clue embeddings. Moreover, it does not address the issue of low-quality or missing clues.

Methods

As shown in Figure 1, we propose a universal model for SS and TSE. The model comprises a backbone network for sound separation, an Encoder-Decoder Based Attractors (EDA) network, and a multi-modal clue network. The EDA network estimates the number of sounds and their attractor embeddings from the mixed audio. The multi-modal clue network takes a variable number of clues with different modalities as input and estimates a unified clue embedding from the clues of different modalities. Align loss is applied between the attractor embeddings and clue embeddings in the training. In the inference stage, the separator can work with the EDA module to perform the SS task; if any clue about the target sound exists, the TSE task can be performed by replacing the attractor embedding with the clue embedding. We will elaborate on the separation network, EDA network, multi-modal clue network, training objectives, and training strategy in the following subsections.

Sound Separation. The sound separation network employs an Encoder-Separator-Decoder structure. Given a single-channel mixed signal $\mathbf{x} \in \mathbb{R}^L$ (where L is of arbitrary length) from J sound sources $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^J$, the network aims to estimate each source signal $\hat{\mathbf{s}}_j$ based on \mathbf{x} and the estimated number of sources \hat{J} from the EDA network.

The encoder uses a 1-D convolution layer with ReLU activation to generate the hidden features $\mathbf{X} \in \mathbb{R}_+^{N \times D}$ of the mixed signal. The separator estimates a mask $\mathbf{M} = \{\mathbf{m}_j\}_{j=1}^J$ for each source. By element-wise multiplication of \mathbf{X} with \mathbf{m}_j , we obtain the estimated hidden representations $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_j\}_{j=1}^J$. Finally, the decoder, composed of transposed convolution layers and symmetric to the encoder, estimates the separated sources $\{\hat{\mathbf{s}}_j \in \mathbb{R}^L\}_{j=1}^J$.

We employed both a time-domain model based on SepFormer (Subakan et al. 2021) and a frequency-domain model based on BSRNN (Luo and Yu 2023) as separators to validate the effectiveness of our joint training framework. Here, we primarily focus on the one based on SepFormer as

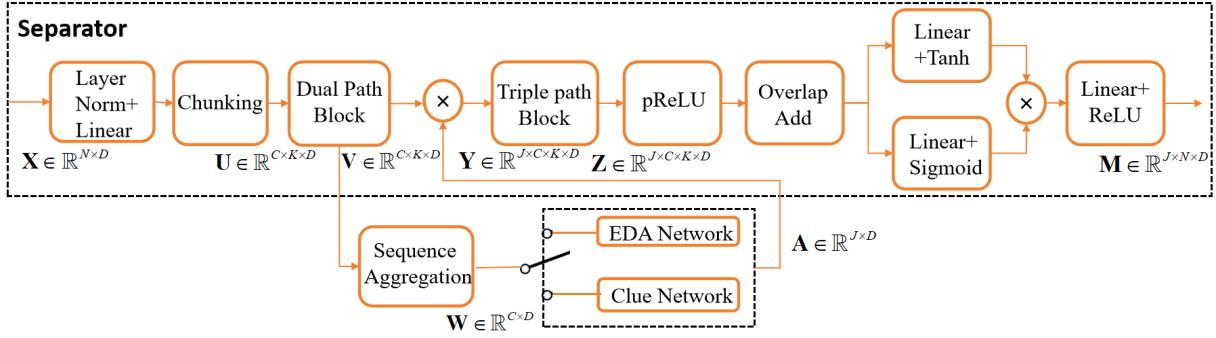


Figure 2: The architecture of the Sepformer-based Separator. It incorporates sequence aggregation to generate \mathbf{W} . For the separation task, it feeds the EDA module to estimate attractors for each sound source in the mixture. For the TSE task, it integrates the Clue Network and features from other modalities for attention computation.

shown in Figure 2. The input feature \mathbf{X} is first processed through a Layer-Norm layer and a linear layer, then segmented into overlapping chunks of size $K = 250$ with a 50% overlap rate. The chunks are fed into a dual-path module, which integrates intra-chunk and inter-chunk transformer blocks (Subakan et al. 2021; Chetupalli and Habets 2022). Sequence aggregation leverages learnable weights to fuse information across distinct feature spaces and the output \mathbf{W} is fed into the EDA block or modality-variant clue network.

The dual-path block’s output \mathbf{V} is element-wise multiplied with the source sound representations $\mathbf{A} \in \mathbb{R}^{J \times D}$ from the EDA or Clue network, forming the input $\mathbf{Y} \in \mathbb{R}^{J \times C \times K \times D}$ for the Triple Path Block. This block extends the dual-path design by adding an inter-channel transformer block, capturing relationships across channels. The final output $\mathbf{Z} \in \mathbb{R}^{J \times C \times K \times D}$ is processed through a Parametric ReLU layer, followed by overlap-add (OVA) and a gated output layer with two linear layers. The final masks m_j are generated via a linear layer with ReLU activation.

EDA Network. The EDA network estimates the number and embeddings of distinct sound categories in mixed audio, enabling sound separation without prior knowledge of source counts. It employs an LSTM Encoder-Decoder framework (Hochreiter 1997) to convert frame-wise embeddings into global attractors.

The LSTM encoder updates its hidden state $\mathbf{h}_t^{\text{enc}}$ and cell state $\mathbf{c}_t^{\text{enc}}$ using the following equations:

$$\mathbf{h}_t^{\text{enc}}, \mathbf{c}_t^{\text{enc}} = h^{\text{enc}}(\mathbf{W}_t, \mathbf{h}_{t-1}^{\text{enc}}, \mathbf{c}_{t-1}^{\text{enc}}) \quad (t = 1, \dots, C). \quad (1)$$

The hidden and cell states of the encoder are initialized to zero vectors: $\mathbf{h}_0^{\text{enc}} = \mathbf{0}$ and $\mathbf{c}_0^{\text{enc}} = \mathbf{0}$.

The LSTM decoder $\mathbf{h}_s^{\text{dec}}$ estimates global-wise attractors as:

$$\mathbf{h}_s^{\text{dec}}, \mathbf{c}_s^{\text{dec}} = h^{\text{dec}}(\mathbf{0}, \mathbf{h}_{s-1}^{\text{dec}}, \mathbf{c}_{s-1}^{\text{dec}}) \quad (s = 1, 2, \dots). \quad (2)$$

At each step, the hidden state $\mathbf{h}_s^{\text{dec}} =: \mathbf{a}_s \in (-1, 1)^D$ serves as the attractor for sound category s , with the dimensionality D matching that of the frame-wise embeddings \mathbf{W}_t . The decoder’s hidden and cell states are initialized by the encoder’s final states: $\mathbf{h}_0^{\text{dec}} = \mathbf{h}_T^{\text{enc}}$ and $\mathbf{c}_0^{\text{dec}} = \mathbf{c}_T^{\text{enc}}$.

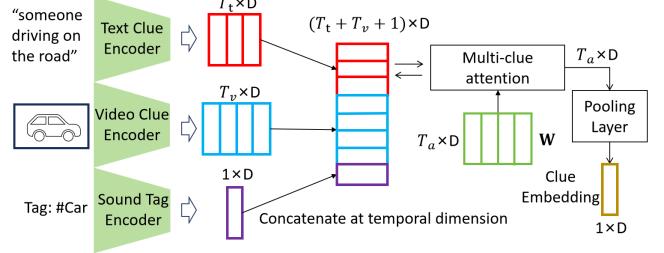


Figure 3: Clue Network Architecture. When any one to three modalities of text, video, and sound tags are present, we employ dedicated pre-trained encoders to extract features from each modality. These features are then concatenated along the temporal dimension and, together with the output \mathbf{W} from sequence aggregation, passed through multi-clue attention and pooling layers to generate the final Clue Embedding.

We compute the attractor existence probabilities using a fully connected layer with a sigmoid activation function, as shown in the following equation:

$$p_{\text{exi}} = \sigma(\mathbf{w}_{\text{exi}}^\top \mathbf{a}_s + b_{\text{exi}}), \quad (3)$$

where $\mathbf{w}_{\text{exi}} \in \mathbb{R}^D$ and $b_{\text{exi}} \in \mathbb{R}$ are the trainable weights and bias parameters of the fully connected layer, respectively. We compare each p_{exi} with a predefined threshold θ in the inference stage. If $p_{\text{exi}} > \theta$, we consider the attractor to exist; otherwise, it is considered that no more sound sources exist in the mixture. The result of the existence check is denoted by q_{exi} .

Clue Network. As shown in Figure 3, the clue network takes variable modality clues (e.g., text, video, sound) as input and generates a comprehensive clue embedding. Each modality is encoded into a unified D dimensional space using dedicated encoders. An attention module then integrates these clues into a temporally aligned fused clue, from which global clue information is derived by averaging over time.

The encoders convert various modalities into D -dimensional embeddings for target sound extraction. The text encoder uses a pre-trained DistilBERT model (Sanh

2019) to transform natural language descriptions into embeddings $\mathbf{O} \in \mathbb{R}^{T_t \times D}$, where T_t is the number of word tokens. The video encoder processes frames with a pre-trained Swin Transformer (Liu et al. 2021), producing video embeddings $\mathbf{V} \in \mathbb{R}^{T_v \times D}$, where T_v is the number of frames. The sound encoder maps one-hot encoded sound event tags to embedding vectors $\mathbf{E} \in \mathbb{R}^{1 \times D}$. All the encoders have their parameters frozen during the training phase.

Multi-Modal Concatenation: The unified multi-modal clue \mathbf{U} is formed by concatenating the text, video, and sound tag embeddings:

$$\mathbf{U} = \text{Concatenate}(\mathbf{O}; \mathbf{V}; \mathbf{E}) \in \mathbb{R}^{(T_t + T_v + 1) \times D}. \quad (4)$$

Attention-Based Clue Fusion: Utilizing the output of Sequence Aggregation \mathbf{W} as the query and \mathbf{U} as both the key and value in the attention mechanism, we derive the fused clue \mathbf{C}_u as follows:

$$\mathbf{C}_u = \text{MultiHeadAttention}(\mathbf{W}, \mathbf{U}, \mathbf{U}) \in \mathbb{R}^{T_a \times D}, \quad (5)$$

where $\text{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes a multi-head attention mechanism (Li et al. 2023; Vaswani et al. 2017) with query \mathbf{Q} , key \mathbf{K} , and value \mathbf{V} . It is evident that the fused clue \mathbf{C}_u shares the same length T_a as the sound embedding \mathbf{Q} .

Loss Function

The objective function is a combination of three distinct losses: the source separation loss \mathcal{L}_{sep} , the source counting loss $\mathcal{L}_{\text{count}}$, and the align loss $\mathcal{L}_{\text{align}}$ between attractors and clues. Specifically, the alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{InfoNCE}}, \quad (6)$$

where \mathcal{L}_{MSE} denotes the Mean Squared Error loss and $\mathcal{L}_{\text{InfoNCE}}$ denotes the InfoNCE loss. Our preliminary experiments showed that using only the MSE Loss lowers attractor–clue alignment accuracy and degrades separation metric SI-SNRi, whereas relying solely on the InfoNCE loss results in slow convergence.

Source Separation Loss (\mathcal{L}_{sep}): In multi-source separation tasks, we apply Signal-to-Noise Ratio (SNR) with Permutation Invariant Training (PIT) as the objective function:

$$\mathcal{L}_{\text{sep}} = - \max_{\pi \in \Pi_K} \sum_{k=1}^K 10 \log_{10} \left(\frac{\sum_{t=1}^T \mathbf{s}_k(t)^2}{\sum_{t=1}^T (\hat{\mathbf{s}}_{\pi(k)}(t) - \mathbf{s}_k(t))^2} \right), \quad (7)$$

where Π_K represents all possible permutations, π is a permutation mapping, K is the number of target sources, T is the length of the signal, $\hat{\mathbf{s}}_{\pi(k)}(t)$ is the value of the estimated source at time step t under permutation π , and $\mathbf{s}_k(t)$ is the value of the target source at time step t .

Source Counting Loss ($\mathcal{L}_{\text{count}}$): This loss is determined using the binary cross-entropy measure, which evaluates the accuracy of the estimated number of sound sources in mixture, given by:

$$\mathcal{L}_{\text{count}} = - \sum_{i=1}^K y_i \log(p_{\text{exi}}) + (1 - y_i) \log(1 - p_{\text{exi}}) \quad (8)$$

where y_i is the true label and p_{exi} is the predicted probability.

MSE Loss (\mathcal{L}_{MSE}): The MSE Loss \mathcal{L}_{MSE} is calculated by first identifying the optimal permutation between the attractors and clues according to final Separation PIT loss. Let π^* denote the best permutation in the Equation (7) and D be the dimension of the embeddings. $\mathbf{a}_{\pi^*(m),i}$ represents the i -th value of the m -th attractor under permutation π^* , while $\mathbf{c}_{m,i}$ is the i -th value of the m -th clue. M is the total number of clues or attractors. The MSE Loss can be expressed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M} \sum_{m=1}^M \frac{1}{D} \sum_{i=1}^D (\mathbf{a}_{\pi^*(m),i} - \mathbf{c}_{m,i})^2 \quad (9)$$

InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$): Given N attractors $\{\mathbf{a}_i\}_{i=1}^N$ and N clues $\{\mathbf{c}_j\}_{j=1}^N$, and the best permutation π^* obtained from the Equation (7), the InfoNCE Loss $\mathcal{L}_{\text{InfoNCE}}$ is computed as the average of the InfoNCE loss (Oord, Li, and Vinyals 2018) for each corresponding pair:

$$\mathcal{L}_{\text{InfoNCE}} = - \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\mathbf{z}_{a_i} \cdot \mathbf{z}_{c_{\pi^*(i)}} / \tau)}{\sum_{j=1}^N \exp(\mathbf{z}_{a_i} \cdot \mathbf{z}_{c_j} / \tau)} \quad (10)$$

where \mathbf{z}_{a_i} is the embedding of the i -th attractor. \mathbf{z}_{c_j} is the embedding of the j -th clue. $\pi^*(i)$ is the index of the clue that is optimally matched to the i -th attractor according to the Separation PIT loss. τ is the temperature parameter.

Training and Inference Strategies

We applied a two-stage training method for model optimization. In stage 1, we train the model for sound separation with an EDA network to estimate the number of sources and compute the source counting loss $\mathcal{L}_{\text{count}}$. Then, the EDA representations are fed into the separation network, which generates separated sounds for calculating the separation loss \mathcal{L}_{sep} . This trains the separation and EDA networks jointly.

In stage 2, we randomly select EDA representations (30% chance) or TSE-generated clue embeddings (70% chance) as inputs to the separation network. For TSE, we train with all seven combinations of modalities (present or absent) equally, using the \mathcal{L}_{sep} loss in a fixed order. We align the two embedding to learn via align loss $\mathcal{L}_{\text{align}}$. This strategy enhances adaptability and robustness.

Inference Phase: In the absence of clues, the EDA module estimates the number and representation of all independent sound sources in the mixture as attractors, which are then fed into the Separator for separation. When one to three modalities of clues are available, the Clue Network generates Clue Embeddings based on these clues, which are subsequently sent into the Separator for targeted extraction.

Experiments

Datasets

To compare our proposed model with existing models, we used the same universal sound dataset as in the (Li et al. 2023) paper, which is based on AudioSet (Gemmeke et al. 2017). AudioSet is a large-scale dataset extracted from YouTube, containing 527 sound classes with weak labels. Each 10-second clip typically includes multiple sound

events without precise timing annotations. To isolate single sound sources, we followed the preprocessing method in (Kong et al. 2020b), using a pre-trained Sound Event Detection (SED) model (Kong et al. 2020a) to identify sound event anchors and extract 2-second audio segments.

We also expanded the 2Mix dataset from (Li et al. 2023) to create a 2&3Mix dataset by remixing existing audio sources, generating 248k mixed samples (140 hours) with 2 or 3 sound sources for training. In the subsequent sections, we will refer to the test sets as Seen datasets. For both the validation and Seen sets, we created 2Mix and 3Mix scenarios. Specifically, the validation sets included 0.5 hours of data, while the Seen datasets consisted of 1 hour of data for each scenario. Additionally, we generated 0.7 hours of data with unseen sound classes (mostly musical instruments) for each scenario to assess the model.

For text clues, we used an audio captioning model (Wu, Dinkel, and Yu 2019) to generate pseudo-natural descriptions from 2-second audio clips. For visual clues, we extracted frames from the aligned 2-second video segments (15 FPS). For tag clues, we converted SED probabilities into one-hot vectors.

Training and Evaluation Details

The experiments were conducted using the ESPnet-SE toolkit (Li et al. 2021). The training was divided into two stages: In stage 1, the learning rate was set to 10^{-4} for 70 epochs, and in stage 2, it was adjusted to 3×10^{-5} for an additional 30 epochs. The threshold θ of EDA is set to 0.5 and the clue embedding dimension D is set to 256.

We used the SNR improvement (SNRi) to evaluate source separation performance. When the estimated number of sources mismatched the ground truth, we adopted the following approach: If the number of sources was overestimated, we retained only the first k estimates for evaluation, where k equals the ground-truth number of sources. If under-estimated, we used a silence signal (all zeros) as the estimate for the missing sources.

Separation Results

Model	Parameters(M)	2 Mix	3 Mix
TDANet-Wav	10.8	17.5	/
TDANet-STFT	7.4	12.7	/
BSRNN-Large	21.8	15.2	/
USE-B	8.1	17.8	15.0
USE-B*	8.1	17.7	15.0

Table 1: Speech separation results on Libri2Mix and our simulated Libri3Mix (SI-SNRi / dB) for USE-B versus competitive models. * represents separation under an unknown and variable number of speakers.

Speech Separation. To assess the ability of the USE framework to perform separation under unknown and variable numbers of speakers, we first compare our unified framework USE-B (based on BSRNN (Luo and Yu 2023)) against the leading universal sound separation models in (Pons et al.

System	2 Mix		3 Mix	
	Seen	Unseen	Seen	Unseen
Sepformer	7.4	6.6	/	/
USE-S(stage 1)	8.7	7.9	6.4	5.2
USE-S(stage 2)	8.8	8.2	7.2	6.3
BSRNN	8.6	8.4	/	/
USE-B(stage 1)	9.2	8.6	5.9	4.8
USE-B(stage 2)	9.4	8.8	6.3	4.8

Table 2: Universal sound separation results for USE-S, USE-B and baseline models (SNRi/dB) on AudioSet dataset, together with ablation studies on EDA Network (stage 1) and Multi-task training strategy (stage 2).

2024) following the similar evaluation adopted in the work on the LibriMix dataset (Cosentino et al. 2020), as shown in Table 1. The results demonstrate that USE-B achieves superior separation performance with a minimal parameter count while remaining robust to both known and unknown numbers of speakers.

Universal Sound Separation. In the universal sound separation experiment, we employed baseline model (Subakan et al. 2021; Luo and Yu 2023) trained on the 2Mix dataset and the separation results are as illustrated in the Table 2. We also assessed the separation performance of models trained using our unified model USE-S (based on Sepformer (Subakan et al. 2021)) and USE-B (based on BSRNN (Luo and Yu 2023)). Our proposed model architecture differs in that it is not restricted to processing a fixed number of audio mixtures. So we trained our USE model on both 2Mix and 3Mix datasets (referred to as stage 1 training) and achieved better SNRi results, representing an improvement over baseline model and thereby validating the superiority of our strategy.

In stage 2 of the training, we integrated TSE with the separation task, simultaneously training for both tasks. We evaluated the separation outcomes of stage 2, and the results showed a slight increase in SNRi. Additionally, the model in stage 2 is also capable of effectively performing TSE task. These findings further substantiate the rationality and effectiveness of our proposed unified training approach for TSE and Sound Separation (SS) tasks.

Evaluation of Source Counting

In our study, we evaluated the accuracy of source count estimation during attractor-based separation. During the inference phase, with the EDA threshold θ set to 0.5, the accuracy of source count estimation can reach over 80% in 2Mix scenario and over 70% in the 3-mix scenario, even on the highly noisy AudioSet dataset. This enables the USE framework to automatically assess and separate sound sources in complex scenarios with unknown numbers of sound sources.

TSE Results

As shown in Table 4, we compared the extraction performance of models trained using our unified framework—such as USE-S and USE-B—against baseline models

Available Clues			DCCRN(Li et al. 2023)		USE-S		USE-B	
tag	text	video	Seen	Unseen	Seen	Unseen	Seen	Unseen
✓	✓	✓	6.9	6.5	8.5	8.1	8.9	8.8
✓	✓		6.8	6.4	8.5	7.9	8.6	8.7
	✓	✓	6.5	6.4	8.2	8.0	8.6	8.7
✓		✓	6.6	6.4	7.8	7.8	8.2	8.4
✓			6.4	6.2	7.3	7.2	7.4	8.0
	✓		6.3	6.0	8.2	7.8	8.0	8.4
		✓	5.8	5.9	6.8	7.2	6.2	7.4

Table 3: TSE Performance comparison of different models with various weakly labelled clues (SNRi/dB).

System	2 Mix		3 Mix	
	Seen	Unseen	Seen	Unseen
Sepformer	8.5	8.0	6.5	5.7
USE-S	8.5	8.1	6.5	5.9
BSRNN	8.4	8.4	5.9	4.8
USE-B	8.9	8.8	6.3	5.0

Table 4: TSE results for USE-S, USE-B and baseline models (SNRi/dB), together with ablation studies on Multi-task training strategy.

Model	2Mix	3Mix
MAE (audio)	5.6	/
USS (audio)	5.6	/
LASS (text)	6.8	/
Audiosep (text)	7.7	/
DCCRN (text+tag+video)	6.9	/
Sepformer (text+tag+video)	8.5	6.5
USE-S (text+tag+video)	8.5	6.5
BSRNN (text+tag+video)	8.4	5.9
USE-B (text+tag+video)	8.9	6.3

Table 5: TSE Performance comparison (SNRi/dB).

trained solely for the TSE (Target Sound Extraction) task. The results show that after joint training with SS and TSE, the extraction performance of USE-S and USE-B did not decline. In fact, there was a slight improvement in TSE performance. Additionally, these models retained the ability to perform separation tasks using the EDA module in the absence of clues.

As shown in Table 3, we compared USE-S and USE-B with DCCRN (Li et al. 2023) using 2Mix data. Our model significantly outperforms DCCRN across all modality configurations, achieving the best performance when all three modalities are used together. The metric of USE-B reaches 8.9 and 8.8 on the Seen and Unseen dataset, respectively. This represents an improvement of 29.0% and 35.4% over DCCRN, respectively. Furthermore, our model also demonstrates strong performance under conditions where only single modality clue is available, thereby proving its superior

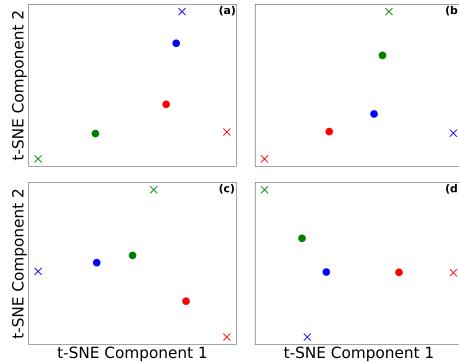


Figure 4: t-SNE visualization of attractors and clues. Different colors denote different sound-source types; circles (\circ) represent attractors, crosses (\times) represent clues. (a) and (b) show the t-SNE visualizations of attractors and clues in the Audioset (Seen) 3-mix, while (c) and (d) show those in the Audioset (Unseen) 3-mix.

robustness and adaptability.

As shown in Table 5, we also compared the performance of different universal sound extraction models on the TSE task. Most of these models used the same SED cropping strategy (Kong et al. 2020b) on the AudioSet dataset as our processing method (Zhao et al. 2024a; Kong et al. 2023; Liu et al. 2022, 2024a; Li et al. 2023; Subakan et al. 2021; Luo and Yu 2023). However, there might be differences in the details, so these results should only be used as a reference. Nevertheless, it is evident from the table that our proposed USE-B model performs the best in TSE task, with performance comparable to that of the separately trained BSRNN+ClueNet model for only TSE task.

Visualization of Attractors and Clues

During the inference process, we evaluated the accuracy of matching clues and attractors by classifying each attractor into one of the clue embeddings via the InfoNCE loss in Equation (10). The accuracy reaches 86.0% in the 2-mix scenario and 65.3% in the 3-mix scenario. This demonstrates that we have effectively unified attractors and clues in the semantic space, enabling them to substitute for each other in the absence of one or the other. This highlights the strong

Datasets	2 Mix		3 Mix		4 Mix		5 Mix		6 Mix	
	FN	PN	FN	PN	FN	PN	FN	PN	FN	PN
Universal-dataset	11.1	11.7	11.3	11.4	10.4	10.3	9.5	9.4	9.0	9.3
Speech+Music	12.3	12.3	9.7	9.3	8.5	8.2	7.1	7.0	/	/
FUSS	14.4	14.7	14.8	15.0	13.1	12.6	/	/	/	/
Audioset(Seen)	8.9	9.2	6.7	6.3	/	/	/	/	/	/
Audioset(Unseen)	8.5	8.8	5.9	5.3	/	/	/	/	/	/

Table 6: Universal Sound Separation Performance of USE-B trained on multi-datasets (SNRi/dB), FN represents Fixed Number of sources, PN represents Predicted Number of sources.

superiority and flexibility of our approach.

In addition, during the inference process, we randomly selected four 3Mix mixtures, two from the Seen dataset and two from the Unseen dataset. We performed t-SNE visualization (Van der Maaten and Hinton 2008) for both attractors and clues in each mixture. As shown in the Figure 4, different sound types have a certain spatial distance in the t-SNE plot, indicating that they possess a certain level of separability in the feature space. In addition, attractors and clues of the same sound type exhibit similarity within the feature space, indicating that they could serve as viable replacements for each other if one is absent.

Universal Sound Separation

We have additionally trained USE-B based on BSRNN on a large-scale general audio dataset, which includes the universal-dataset (2~6 mix) mainly remixed by VGGSound (Chen et al. 2020), AudioSet (Gemmeke et al. 2017) (2~3 mix), FUSS (Wisdom et al. 2021) (2~4 mix), and Mu-san (Snyder, Chen, and Povey 2015) (music) + Librispeech (Panayotov et al. 2015) (speech) (2~5 mix). During the evaluation, we considered two situations: one where the number of sources to be separated is known in advance, and another where the number of sources is unknown and needs to be estimated using the EDA module. The evaluation results are as Table 6. Firstly, we can see that USE-B is capable of adapting to mixtures with varying numbers of sources and performs well across different datasets. Secondly, we have found that the method of predicting the number of sound sources using the EDA module generally matches or outperforms the approach based on a pre-determined fixed number of sources.

We also compared our USE-B model with other competitive models on the FUSS (Wisdom et al. 2021; Wang et al. 2021)(Dry, without reverberation) test set, as shown in Table 7. It can be observed that our model achieved superior separation performance in both 2-mix, 3-mix, and especially in 4-mix scenarios.

We have also evaluated the inference speed of extracting 1 to 6 source audio signals from a mixture as shown in Figure 5. It is observed that the computational complexity during inference increases linearly with the number of sound sources. Notably, even when inferring up to six sources, the computational demand in terms of GFLOPS remains below 30. This characteristic ensures the real-time performance and high efficiency of the USE-B model during inference.

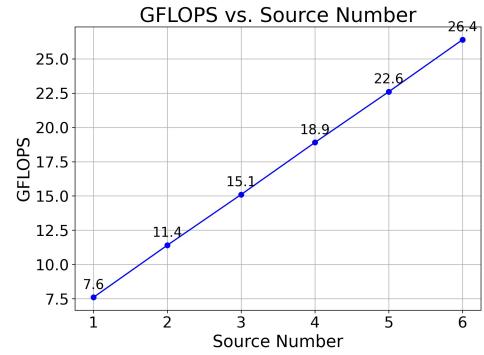


Figure 5: The GFLOPS of the USE Model as the number of sources increases.

Model	2 Mix	3 Mix	4 Mix
TDCN++	11.2	11.6	7.4
USE-B	12.8	13.1	11.9

Table 7: Universal Sound Separation Performance Comparison on FUSS (Dry) dataset (SI-SNRi/dB).

Conclusion

In this study, we introduced USE, a universal method that unifies the Sound Separation and Target Sound Extraction tasks. This novel approach can handle diverse sound types, variable numbers of sources, and multiple modalities of clues and outperforms existing models in multi-source separation and target sound extraction tasks thanks to its ability to integrate an arbitrary number of clues. However, we found that the performance of USE can be influenced by the cleanliness of individual sound sources in the training data. One promising direction is to develop methods for dynamically adjusting the granularity of attractors to improve its adaptability to various tasks. In different acoustic scenarios, some sound sources need to be separated individually, while others should be grouped together. Additionally, integrating USE with sound event detection and understanding tasks could provide a more comprehensive approach to audio processing. By combining these capabilities, we can develop a more robust framework and offer deeper insights into the acoustic environment.

Acknowledgements

This work was supported in part by China STI 2030-Major Projects under Grant No. 2021ZD0201500, and in part by China NSFC project under Grants No. U25A20409.

References

- Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vgssound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.
- Cheng, X.; Zheng, S.; Wang, Z.; Fang, M.; Zhang, Z.; Huang, R.; Ma, Z.; Ji, S.; Zuo, J.; Jin, T.; et al. 2024. OmniSep: Unified Omni-Modality Sound Separation with Query-Mixup. *arXiv preprint arXiv:2410.21269*.
- Chetupalli, S. R.; and Habets, E. A. 2022. Speech Separation for an Unknown Number of Speakers Using Transformers With Encoder-Decoder Attractors. In *INTERSPEECH*, 5393–5397.
- Chetupalli, S. R.; and Habets, E. A. 2024. A Unified Approach to Speaker Separation and Target Speaker Extraction Using Encoder-Decoder Based Attractors. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 190–194. IEEE.
- Cosentino, J.; Pariente, M.; Cornell, S.; Deleforge, A.; and Vincent, E. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.
- Delcroix, M.; Ochiai, T.; Zmolikova, K.; Kinoshita, K.; Tawara, N.; Nakatani, T.; and Araki, S. 2020. Improving speaker discrimination of target speech extraction with time-domain speakerbeam. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 691–695. IEEE.
- Elminshawi, M.; Mack, W.; Chetupalli, S. R.; Chakrabarty, S.; and Habets, E. A. 2022. New insights on target speaker extraction. *arXiv preprint arXiv:2202.00733*.
- Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.
- Hochreiter, S. 1997. Long Short-term Memory. *Neural Computation MIT-Press*.
- Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K.; Le Roux, J.; and Hershey, J. R. 2019. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 175–179. IEEE.
- Kilgour, K.; Gfeller, B.; Huang, Q.; Jansen, A.; Wisdom, S.; and Tagliasacchi, M. 2022. Text-driven separation of arbitrary sounds. *arXiv preprint arXiv:2204.05738*.
- Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumley, M. D. 2020a. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.
- Kong, Q.; Chen, K.; Liu, H.; Du, X.; Berg-Kirkpatrick, T.; Dubnov, S.; and Plumley, M. D. 2023. Universal source separation with weakly labelled data. *arXiv preprint arXiv:2305.07447*.
- Kong, Q.; Wang, Y.; Song, X.; Cao, Y.; Wang, W.; and Plumley, M. D. 2020b. Source separation with weakly labelled data: An approach to computational auditory scene analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 101–105. IEEE.
- Lee, Y.; Choi, S.; Kim, B.-Y.; Wang, Z.-Q.; and Watanabe, S. 2024. Boosting unknown-number speaker separation with transformer decoder-based attractor. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 446–450. IEEE.
- Li, C.; Qian, Y.; Chen, Z.; Wang, D.; Yoshioka, T.; Liu, S.; Qian, Y.; and Zeng, M. 2023. Target sound extraction with variable cross-modality clues. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Li, C.; Shi, J.; Zhang, W.; Subramanian, A. S.; Chang, X.; Kamo, N.; Hira, M.; Hayashi, T.; Boeddeker, C.; Chen, Z.; et al. 2021. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 785–792. IEEE.
- Liu, X.; Kong, Q.; Zhao, Y.; Liu, H.; Yuan, Y.; Liu, Y.; Xia, R.; Wang, Y.; Plumley, M. D.; and Wang, W. 2024a. Separate anything you describe. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Liu, X.; Liu, H.; Kong, Q.; Mei, X.; Zhao, J.; Huang, Q.; Plumley, M. D.; and Wang, W. 2022. Separate what you describe: Language-queried audio source separation. *arXiv preprint arXiv:2203.15147*.
- Liu, Y.; Liu, X.; Zhao, Y.; Wang, Y.; Xia, R.; Tain, P.; and Wang, Y. 2024b. Audio Prompt Tuning for Universal Sound Separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1446–1450. IEEE.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Luo, Y.; and Mesgarani, N. 2019. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8): 1256–1266.
- Luo, Y.; and Yu, J. 2023. Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1893–1901.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio

- books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.
- Pons, J.; Liu, X.; Pascual, S.; and Serrà, J. 2024. Gass: Generalizing audio source separation with large-scale data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 546–550. IEEE.
- Saijo, K.; Zhang, W.; Wang, Z.-Q.; Watanabe, S.; Kobayashi, T.; and Ogawa, T. 2023. A Single Speech Enhancement Model Unifying Dereverberation, Denoising, Speaker Counting, Separation, and Extraction. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–6. IEEE.
- Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Snyder, D.; Chen, G.; and Povey, D. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.
- Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; and Zhong, J. 2021. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 21–25. IEEE.
- Takahashi, N.; Parthasarathy, S.; Goswami, N.; and Mitsufuji, Y. 2019. Recursive speech separation for unknown number of speakers. *arXiv preprint arXiv:1904.03065*.
- Tzinis, E.; Wang, Z.; Jiang, X.; and Smaragdis, P. 2022. Compute and memory efficient universal sound source separation. *Journal of Signal Processing Systems*, 94(2): 245–259.
- Tzinis, E.; Wisdom, S.; Hershey, J. R.; Jansen, A.; and Ellis, D. P. 2020. Improving universal sound separation using sound classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 96–100. IEEE.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- Wang, Z.-Q.; Erdogan, H.; Wisdom, S.; Wilson, K.; Raj, D.; Watanabe, S.; Chen, Z.; and Hershey, J. R. 2021. Sequential multi-frame neural beamforming for speech separation and enhancement. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 905–911. IEEE.
- Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.
- Wisdom, S.; Erdogan, H.; Ellis, D. P.; Serizel, R.; Turpault, N.; Fonseca, E.; Salamon, J.; Seetharaman, P.; and Hershey, J. R. 2021. What’s all the fuss about free universal sound separation data? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 186–190. IEEE.
- Wu, M.; Dinkel, H.; and Yu, K. 2019. Audio caption: Listen and tell. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 830–834. IEEE.
- Zhang, K.; Li, J.; Wang, S.; Wei, Y.; Wang, Y.; Wang, Y.; and Li, H. 2024. Multi-Level Speaker Representation for Target Speaker Extraction. *arXiv preprint arXiv:2410.16059*.
- Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, 570–586.
- Zhao, J.; Liu, X.; Zhao, J.; Yuan, Y.; Kong, Q.; Plumley, M. D.; and Wang, W. 2024a. Universal sound separation with self-supervised audio masked autoencoder. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, 1–5. IEEE.
- Zhao, S.; Ma, Y.; Ni, C.; Zhang, C.; Wang, H.; Nguyen, T. H.; Zhou, K.; Yip, J. Q.; Ng, D.; and Ma, B. 2024b. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10356–10360. IEEE.