Journal of *Marine Science and Engineering*

MDPI

Article

A Multi-Task Network: Improving Unmanned Underwater Vehicle Self-Noise Separation via Sound Event Recognition

Wentao Shi [1,2], Dong Chen [2], Fenghua Tian [3], Shuxun Liu [2] and Lianyou Jing [1,*]

[1] Ocean Institute, Northwestern Polytechnical University, Taicang 215400, China; swt@nwpu.edu.cn
[2] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; chendong0690@mail.nwpu.edu.cn (D.C.); shuxunliu@mail.nwpu.edu.cn (S.L.)
[3] Xi'an Precision Machinery Research Institute, Xi'an 710077, China; tfh2000526@163.com
* Correspondence: lyjing@nwpu.edu.cn

Abstract: The performance of an Unmanned Underwater Vehicle (UUV) is significantly influenced by the magnitude of self-generated noise, making it a crucial factor in advancing acoustic load technologies. Effective noise management, through the identification and separation of various self-noise types, is essential for enhancing a UUV's reception capabilities. This paper concentrates on the development of UUV self-noise separation techniques, with a particular emphasis on feature extraction and separation in multi-task learning environments. We introduce an enhancement module designed to leverage noise categorization for improved network efficiency. Furthermore, we propose a neural network-based multi-task framework for the identification and separation of self-noise, the efficacy of which is substantiated by experimental trials conducted in a lake setting. The results demonstrate that our network outperforms the Conv-tasnet baseline, achieving a 0.99 dB increase in Signal-to-Interference-plus-Noise Ratio (SINR) and a 0.05 enhancement in the recognized energy ratio.

Keywords: self-noise; identification and separation; sound event recognition; neural network; multi-task network

Citation: Shi, W.; Chen, D.; Tian, F.; Liu, S.; Jing, L. A Multi-Task Network: Improving Unmanned Underwater Vehicle Self-Noise Separation via Sound Event Recognition. *J. Mar. Sci. Eng.* **2024**, *12*, 1563. https://doi.org/10.3390/jmse12091563

Academic Editor: Leszek Chybowski

Received: 20 July 2024
Revised: 28 August 2024
Accepted: 4 September 2024
Published: 5 September 2024

1. Introduction

Covering over 70% of the earth's surface, the oceans are vast repositories of mineral and biological wealth, within which UUVs play a pivotal role in exploration and exploitation endeavors [1–4]. In the realms of UUV development and maintenance, engineers frequently rely on hydrophones to detect and analyze the vehicle's acoustic emissions, known as UUV self-noise, to assess its condition [5,6]. These acoustic data are instrumental for downstream applications such as noise control and fault diagnosis [7–9]. However, the challenge lies in the composite nature of UUV self-noise, which includes flow noise, vibration noise, and electrical noise. These components can interfere with specific analytical tasks. For instance, when attempting to mitigate one type of noise's impact on the Sound Pressure Level (SPL) across various frequency bands, the remaining noise types may prove detrimental [10,11]. Similarly, vibration noise components are particularly critical for fault detection and diagnosis tasks in UUVs [11]. Addressing these challenges necessitates an innovative approach: the separation of UUV self-noise into its constituent single-source components, a process akin to sound source separation. This technique promises to enhance the precision of UUV diagnostics and operational efficiency by isolating and analyzing the distinct noise types.

Recent advancements in Deep Neural Networks (DNNs) have significantly advanced the field of sound source separation, marking a notable improvement in the extraction of individual audio components from complex signal mixtures [12–14]. This capability positions sound source separation as an invaluable pre-processing tool for a wide array of signal processing tasks [15–18]. Particularly in the context of self-noise analysis for

UUVs, the ability to isolate single sources from a blend of noises can lead to substantial enhancements in data quality and analysis accuracy. Therefore, the adoption of deep learning-based techniques for sound source separation in the analysis of UUV self-noise is poised to offer considerable benefits, enriching the quality of downstream processing tasks and contributing to more refined and effective UUV operational strategies.

To realize the objectives outlined previously, it is essential to first delineate the specific aims of UUV self-noise separation and ascertain any requisite customizations. In adapting sound source separation techniques to the UUV context, we must account for three principal distinctions relative to their application in other domains: data volume, data variability, and separation objectives. The volume of available data is a pivotal consideration, directly influencing the neural network's ability to achieve convergence. Underwater environments, characterized by numerous uncontrollable variables, coupled with the high costs associated with underwater experimentation, render the acquisition of isolated source noises from UUVs both challenging and expensive. Moreover, the data collected from UUV operations under ostensibly similar conditions exhibit minimal intrinsic variability. This consistency in operational data necessitates a nuanced approach to network design to effectively discern and isolate noise sources. Finally, the aim of UUV self-noise separation extends beyond the simple extraction of individual sound components. A critical, and often neglected, aspect within this specific field is the preservation of the original energetic proportions of each noise source. This requirement presents a unique challenge in the UUV self-noise domain, underscoring the need for a tailored approach to sound source separation that transcends conventional waveform recovery.

In this study, our aim is to synthesize the strengths of various methodologies, some of which have been preliminarily applied to UUV self-noise, to devise a solution that more closely aligns with our objectives. The majority of extant deep learning-based research on source separation has predominantly focused on speech, music, and bioacoustics [19–24]. A smaller contingent of scholars has explored the separation of arbitrary sound sources [25–27]. Notably, the underlying strategies employed across these domains bear significant similarities, primarily revolving around two main types of training objectives: mapping-based and mask-based targets [28]. Drawing from historical insights and our empirical findings, we consistently observe superior outcomes with mask-based approaches over mapping-based strategies [29]. To address the unique challenges inherent in UUV self-noise separation, we propose leveraging multi-task learning. This approach facilitates concurrent separation and recognition, offering a dual advantage. Consequently, we introduce a preliminary framework for our network: a mask-based, multi-task learning architecture specifically designed for the nuanced requirements of UUV self-noise analysis.

While a preliminary concept lays the groundwork, the integration of cutting-edge technologies is imperative to refine our approach. Variational Autoencoders (VAEs) stand out in the realm of self-supervised learning systems, characterized by their capacity to approximate the input as the output [30]. Renowned for their efficacy in generative modeling, VAEs adeptly capture complex data distributions through a framework rooted in the traditional autoencoder architecture [31,32]. This structure comprises an encoder, which distills the input into a condensed latent representation, and a decoder, which reconstructs the input from this latent space. Distinctively, VAEs incorporate probabilistic elements—specifically, a regularization term—that enable them not only to derive meaningful representations but also to generate novel data samples. In light of the structural and interpretive advantages offered by VAEs, we propose a VAE-inspired approach tailored to UUV self-noise analysis. Our model employs an encoder to deduce latent representations of noise masks and a decoder to approximate these masks. A novel aspect of our methodology is the integration of UUV navigational conditions into the VAE's probabilistic regularization framework. The sound events, readily extractable from audio data, provide contextual cues that enhance the model's performance across tasks. Specifically, for UUV self-noise separation, the navigational phases—innate characteristics of the self-noise—act as a dynamic regularization element, enhancing the model's adaptability. We introduce a 'sound

event enhanced module' designed to augment separation accuracy while concurrently recognizing energetic ratios. This module interprets the probabilistic regularization term as a latent or generalized variable, facilitating the transformation of mask data in the latent space into more interpretable latent variables, thereby generalizing the separated waveform with reduced reconstruction error. This process constrains the latent space to a more defined and interpretable region. Although inspired by traditional VAEs, our model diverges in its focus on mask representations rather than the original data's representations. Furthermore, unlike VAEs, which primarily function as generative models for specific signal synthesis, our model is not designed for signal generation per se. This distinction underscores the tailored application of VAE principles to the unique challenges of UUV self-noise separation.

In this study, we introduce a lightweight, multi-task network specifically designed to address the challenges of UUV self-noise separation. The remainder of this paper is structured as follows: Section 2 reviews the pertinent literature, laying the groundwork for our approach. Section 3 delineates the architecture and functionalities of the Multi-Task Learning Network (MLN). Section 4 details the experimental setup and parameters. Finally, Sections 5 and 6 discuss our findings and draw conclusions, respectively, highlighting the implications and potential future directions of our research.

2. Related Work

Sound source separation, often termed blind source separation, has garnered substantial interest within the signal processing community for its ability to discern individual audio signals from composite mixtures without prior knowledge of the source processes. The quintessential 'cocktail party problem' epitomizes the core challenge of this field, illustrating how a listener might focus on a single conversation amidst a cacophony of background noise, akin to isolating specific sound sources from a complex auditory environment. Initially, methodologies such as Independent Component Analysis (ICA) [33,34] and Non-negative Matrix Factorization (NMF) [35,36] were employed to decompose audio mixtures into discrete time-domain waveforms. However, the advent of deep learning has broadened the scope of sound source separation, leading to its application in specialized subdomains like speech, music, and bioacoustics separation [12,15,29]. This paper delineates deep learning-based sound source separation strategies into two primary categories: traditional data-driven approaches and those utilizing VAE techniques [32,37]. The former relies on extensive datasets to train models capable of identifying and isolating sound sources, while the latter employs VAE frameworks to achieve separation by leveraging probabilistic modeling and generative capabilities.

Over the past decade, data-driven approaches to sound source separation have witnessed remarkable advancements. Huang et al. were the first to integrate the traditional masking mechanism with deep learning for sound source separation [38]. Hershey et al. innovatively merged deep learning networks with k-means clustering, facilitating the separation of distinct sources through the clustering of deep embeddings [39]. Yu et al. advanced this approach by optimizing deep clustering for end-to-end training in the time–frequency (T-F) domain [40], while Wang et al. leveraged deep clustering as a regularization term, thereby enhancing separation performance [29]. Kolbak et al. introduced utterance-level permutation invariant training, a novel strategy aimed at resolving the permutation ambiguity in the separation process of different sources [41]. Luo et al. developed the Audio Separation Network, which processes signals in the time domain and surpasses previous methods that operated in the T-F domain in terms of performance [19]. Pishdadian et al. adopted class labels as proxies for signal-level ground truth, implementing a weakly supervised approach to separation [42]. Seetharaman et al. exploited conditional embeddings derived from each source class within a Gaussian kernel space, significantly boosting separation efficacy [43].

Since its inception, the VAE has been widely adopted in a plethora of domains, including image classification, generation, denoising, and anomaly detection, and has made

significant strides in the realm of sound source separation. Do et al. leveraged VAEs by inputting time–frequency (T-F) spectrograms of mixed signals, coupled with a Chebyshev bandpass filter on the output, to facilitate source separation [37]. Rather than relying on the source signals directly, one approach utilized class information to drive the separation process, employing a β-VAE with convolutional layers for this task [44]. The study by Grais et al. utilized convolutional denoising autoencoders, taking the magnitude spectrogram of the mixed signal as input for monoaural source separation [45]. In the context of the Voice Conversion Challenge, the pursuit of semi-blind source separation led to the development of the Multichannel Conditional VAE (MCVAE) method, which demonstrated commendable performance under underdetermined conditions [46]. Despite the generalized MCVAE's effectiveness in source separation, it is hampered by computational complexity and does not achieve high accuracy in source classification, prompting the introduction of the fast MCVAE to address these limitations [47]. While existing VAE-based algorithms exhibit robust capabilities, directly applying these methods to UUV self-noise separation may not produce optimal results. This highlights the need for tailored approaches that account for the unique characteristics and challenges associated with UUV self-noise.

3. Underwater Acoustic Multi-Task Noise Separation and Recognition Method

This section begins with an overview of fundamental concepts pertinent to the mask-based single-channel source separation challenge. Subsequently, we introduce our novel 'Sound Event Enhanced Separation Module', delineate the formulation of the objective function, and elucidate the training methodology in detail.

3.1. Mask-Based Single-Channel Noise Source Separation

The normal single-channel noise separation can be formulated into estimating C sources $s_1(t), ..., s_C(t) \in \mathbb{R}^{1\times T}$ from given mixture $x(t) \in \mathbb{R}^{1\times T}$, where

$$x(t) = \sum_{i=1}^{C} s_i(t). \tag{1}$$

Generally, before processing the entire UUV self-noise, the mixed noise and the noise from each clean source are first segmented into independent noise sample segments to prepare for subsequent processing steps. Each vector can be represented by a set of base vectors under it, that is, each segment of mixed noise and clean noise can be represented by the product of a weight matrix and a base vector:

$$\begin{cases} x = v \odot \boldsymbol{B} \\ s_c = w_c \odot \boldsymbol{B} \end{cases} \tag{2}$$

where $\boldsymbol{B} = [b_1, b_2 \cdots, b_N]$ is the basis vector, $v \in \mathbb{R}^{1\times N}$ is the weight vector matrix of the mixed noise, and $w_c \in \mathbb{R}^{1\times N}$ is the weight vector matrix of the c-th noise source. The mixed noise is the sum of noise source in the time domain. When the basis vectors remain consistent, the weight matrix of mixed noise can be obtained by adding the weight matrices of various noise sources. Th relationship between these two weight matrices is as follows:

$$X = \sum_{c=1}^{C} w_c \tag{3}$$

It can be seen that when the result of multiplying the weight matrix with the basis vector is used to represent mixed noise and clean noise, the noise separation problem can be transformed into a convolutional operation form. Considering the non-negativity of the weight matrix w_c, the separation of mixed noise can be seen as estimating the encoder output and constructing a mask m_c for separating noise based on this. By multiplying the

mask m_c with the mixed noise weight matrix, the weight matrices w_c of each noise source can be identified one by one

$$w_c = \mathrm{m}_c \odot v \tag{4}$$

where $\odot$ represents the dot product.

In the mask-based separation mechanism, there are three main blocks, the encoder, mask-estimator, and decoder, and each of them plays a unique role. In an encoder, the mixture waveform $x(t)$ will be mapped to a high-dimensional representation $\mathrm{w} \in \mathbb{R}^{N \times T}$, which could be reformulated as a matrix multiplication:

$$\mathrm{w} = \mathcal{H}(x\mathrm{U}^{\mathrm{T}}), \tag{5}$$

where $\mathrm{U} \in \mathbb{R}^{N \times T}$ includes N vectors with length T each and $\mathcal{H}(\cdot)$ represents the rectified linear unit (ReLU) function in [38]. The mask-estimator will generate C masks $\mathrm{m_i} \in \mathbb{R}^{1 \times N}$, $i = 1, \cdots, C$ where C is the number of the noise type in the mixture and $\mathrm{m_i}$ is the mask vector which has the constraint that $\mathrm{m_i} \in [0, 1]$ and $\sum_{i=1}^{C} \mathrm{m}_i = 1$. Through a simple element-wise multiplication, the representation of each source, $\mathrm{d}_i \in \mathbb{R}^{1 \times T}$, can be calculated as follows:

$$\mathrm{d}_i = \mathrm{w} \odot \mathrm{m}_i. \tag{6}$$

Finally, a decoder will reconstruct the original waveform of each source to the estimated source $\hat{\mathrm{s}}_i(t) \in \mathbb{R}^{1 \times T}$, $i = 1, \ldots, C$, as follows:

$$\hat{\mathrm{s}}_i = \mathrm{d}_i \mathrm{V}, \tag{7}$$

where $\mathrm{V} \in \mathbb{R}^{N \times T}$ is the matrix with the rows in the decoder basis functions with length T.

In the three blocks we mentioned above, the encoder and the decoder are achieved by regular convolutional and transposed convolutional layers which can enable better convergence and faster training. The mask-estimator is introduced with our proposed event label improved separation module in the next subsection. We briefly show the mask-based separation workflow with the improved module in Figure 1.

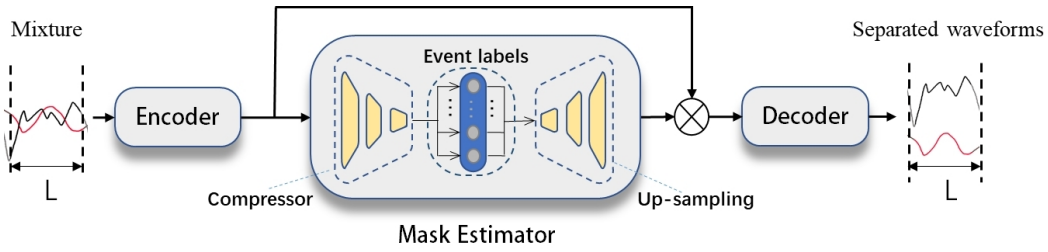


Figure 1. Mask-based separation mechanism diagram.

3.2. The Mask-Estimator with Sound Event Improved Module

At a foundational level, the mask-estimator is segmented into two primary functions: classification and up-sampling. The classification component is tasked with refining the network's representation w into more succinct features z and determining the energetic ratios. Conversely, the up-sampling function projects this condensed representation z onto the mask m for disentanglement. Within our framework, the classification phase is further decomposed into a compressor and an enhancement module. The compressor focuses on the reduction in data dimensionality, while the enhancement module, dubbed the 'Sound Event Improved Module', extends the capabilities of the classifier. Positioned between the compressor and the up-sampling stages, this module processes the compressed representation z, enriches it with information from the sound event labels, and forwards the enriched representation for up-sampling. Notably, the compressor and up-sampling

components mirror each other in structure but are inverse in function. A more detailed exposition of these network elements and their interactions will be presented in Section 3.4.

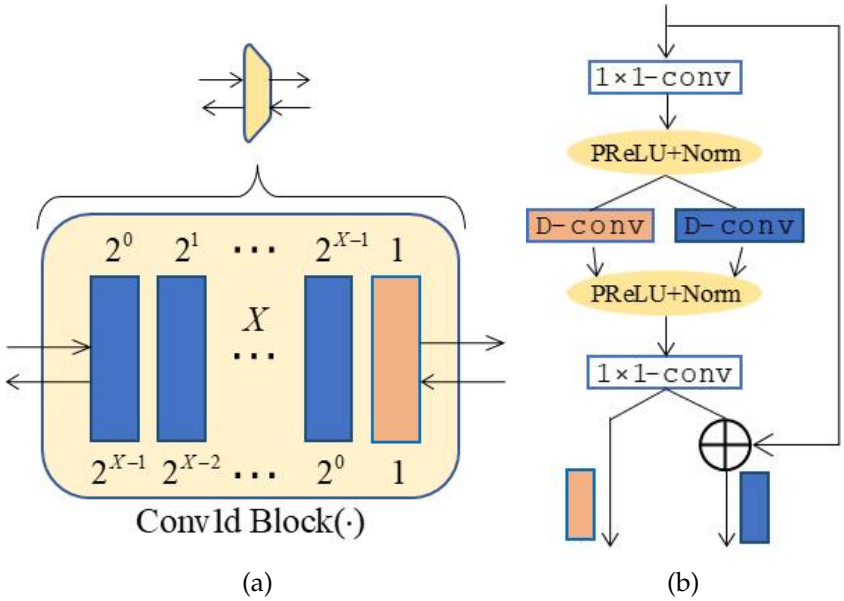
3.2.1. The Compressor and Up-Sampling

The main components in the mask-estimator is the chunk which is composed of several convolutional blocks (Conv1d Block$(\cdot)$) with increasing dilation factors, as denoted in Figure 2a. In the compression process, the representations are propagated through X convolutional blocks from left to right, with the dilation factors growing exponentially. Conversely, in the up-sampling phase, the progression follows the reverse direction.

Figure 2b shows the design of each single convolutional block where two outputs are applied: a residual path in the block serves as the exclusive part of the blue block, and a forward pass with dimensional change over time is only available in the orange block. To decrease the number of parameters, we replace the standard convolution with the depthwise separable convolution that comprises pointwise convolution $1 \times 1 - conv$ and depthwise convolution $D - conv$ in series. The depthwise convolution in blue adopts the dilation factors shown in Figure 2a, while the orange executes with no dilation factors.

A yellow trapezoid visualizes the change in the length of representation before and following the process. In the compressor, the representation ingresses from the left and egresses from the right, resulting in a lower dimension of time, while the opposite occurs in the up-sampling. The blue blocks share the same structures and maintain the same time dimensions, while the orange block reduces dimensionality in the compressor and increases dimensionality in the up-sampling. X refers to the total number of the blue blocks and the order of two indicates the dilation factors.

The forward propagation is represented by the single block in Figure 2a, where the different outputs according to the position in Figure 2a are illustrated by colors. Each block is composed of two $Conv1 \times 1$ operation and a conditional depthwise convolution ($D - conv$), with PreLU activation function and global layer normalization situated between each two convolution operations. There are two outputs of the block, a direct output represented by the orange block and a residual path output through the blue block. The conditional depthwise convolution exhibits different functions corresponding to the color of the outputs.

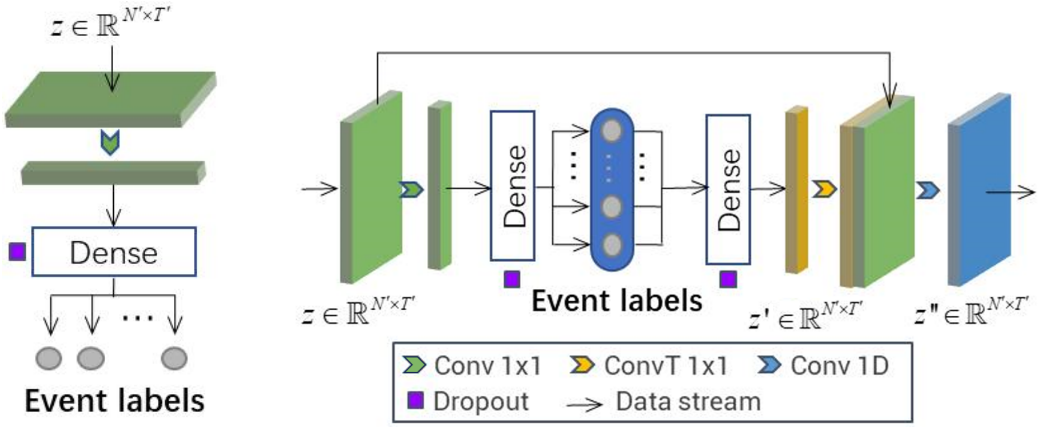


(a) (b)

Figure 2. The Convld block design in the compressor and the up-sampling: (**a**) Convld block; (**b**) Single convolutional block.

3.2.2. Sound Event Improved Module

Before delving into the specifics of the Sound Event Improved Module, it is pertinent to outline two distinct frameworks that both align with our objectives. These frameworks are depicted in Figure 3. Figure 3a presents the DV, which solely focuses on recognition outcomes without engaging in latent space regularization. In contrast, Figure 3b illustrates the Standard Version (SV), which not only aims for recognition, but also enhances separation performance through meticulous regularization of the latent space. The intricacies and functionalities of these two modules will be elaborated upon in the subsequent subsection.



(a) (b)

Figure 3. The two kinds of frameworks of the Sound Event Improved Module diagrams: (**a**) The DV without the function of regularization; (**b**) The SV, which outputs both the recognition results and regularization terms.

(a) The DV Module

For the question related to the multi-tasks, the DV module is a straightforward solution that recognizes the energetic ratio through a simple class module. In detail, DV acts as a classifier and is described below:

$$P = MLP(Conv1 \times 1(z)), \tag{8}$$

where $P = p_1(t), \ldots, p_C(t) \in \mathbb{R}^C$ is the estimated energetic ratio, $Convl \times 1(\cdot)$ and $MLP(\cdot)$ represent the operation of a one-dimension convolutional layer with the kernel of one and a dense layer. In this no-regularization case, the estimated results identified by the DV should only match the correct sound event labels. This can be accomplished by incorporating a binary cross-entropy term between the true labels and the corresponding outputs. Let $H(l, p)$ denote the binary cross-entropy function, which is defined as

$$H(l, p) = -l \log(p) - (1 - l) \log(1 - p), \tag{9}$$

where $l \in [0, 1]$ denotes the true class probabilities while $p \in [0, 1]$ is the estimated. We denote by $\mathcal{L}_{class}(x, L)$ the energetic recognition loss for an input waveform $x(t)$ and its associated energetic ratio labels $L = \{l_1, l_2, \ldots, l_C\}$, where $l_1(t) + l_2(t) \ldots + l_C(t) = 1$. We also restrict the sum of the estimated energetic ratio P to be equal to one. For the mixture $x(t)$, the recognition loss can be empirically computed by aggregating the sum of binary cross-entropy terms over each individual source:

$$\mathcal{L}_{class}(x, L) = \sum_{i=1}^{C} H(l_i, p_i(x)). \tag{10}$$

(b) The SV Module

The SV is an evolutionary variant of the DV. In the SV, we seek not only the results of the recognition and separation but, more significantly, their mutual reinforcement. In other words, the output in the DV is used to be the latent variables and improve the performance of separation in the SV. As illustrated in Figure 3b, the SV structure is similar to a mirror, with the output region (blue) as the axis of symmetry. When receiving the representation z from the compressor, the SV executes the same operation as the DV to obtain the latent variables (i.e., the energetic ratio), regularizes the variables with the true event labels by cross-entropy loss function, and later, recovers the regularized representation z', which is of the same shape as z. There is also a skip connection between z' and z. The formulation of the additional module in the SV is as follows:

$$z' = Conv1 \times 1(MLP(P)). \tag{11}$$

$$z'' = Conv1d(concat(z, z')), \tag{12}$$

where z'' is the input of the up-sampling generated by the Sound Event Improved Module and $Convld(\cdot)$ and $concat(\cdot)$, respectively, represent a one-dimensional convolution layer and the concatenation of two characteristic maps.

3.3. Balancing Multi-Task Weights

In our discussion so far, we have framed the multi-task learning paradigm within the context of simultaneous recognition and separation tasks. This paradigm inherently lends itself to a multi-objective optimization framework, wherein we seek to balance and optimize a set of potentially competing objectives. For the purposes of this study, we focus on a dual-task scenario. The loss function associated with the recognition task is presented in Equation (10), while the objective for the separation task is to maximize the Scale-Invariant Source-to-Noise Ratio (SISNR), a metric defined as follows:

$$\text{SISNR}(\mathrm{s}, \hat{\mathrm{s}}) = 10 \log_{10} \frac{\| \alpha \mathrm{s} \|^2}{\| \alpha \mathrm{s} - \hat{\mathrm{s}} \|^2}, \tag{13}$$

$$\mathcal{L}_{sep} = -\sum_{c=1}^{C} (\text{SISNR})(s_c, \hat{s}_c), \tag{14}$$

where $\hat{\mathrm{s}} \in \mathbb{R}^{1 \times T}$ and $\mathrm{s} \in \mathbb{R}^{1 \times T}$ are the estimated and original clean sources, respectively; $\alpha = \hat{s}\mathrm{s}^T / \|\mathrm{s}\|^2$; $\|\mathrm{s}\|^2 = < \mathrm{s}, \mathrm{s} >$ is the signal power; and $< \cdot, \cdot >$ is the scalar product. Permutation invariant training (PIT) is implemented during the training phase to address the source permutation problem [40,41]. The total loss in our network is thus composed of $\mathcal{L}_{class}$ and $\mathcal{L}_{sep}$ with a computed weight β defined as follows:

$$L_{total} = \beta L_{class} + (1 - \beta)\mathcal{L}_{sep}. \tag{15}$$

It is customary to determine the value of β either through an expensive grid search across various scaling or through the application of a heuristic approach. There is a better treatment that if the network is able to update β in the course of training, it will be more likely to acquire global optimality. The updating scheme of hyper-parameter β will be provided with the updated algorithm of the network. The loss functions of the network are shown in Table 1.

Table 1. Summary of the loss function.

Task Type	Equation	Loss Function
Recognition	(10)	$\mathcal{L}_{class}$
Separation	(14)	$\mathcal{L}_{sep} = -\sum_{c=1}^{C} (\text{SISNR})(s_c, \hat{s}_c)$
Total	(15)	$L_{total} = \beta L_{class} + (1-\beta)\mathcal{L}_{sep}$

3.4. Network Architecture

The overall structure of our network can be divided into three parts, encoder, decoder and mask-estimator, whereby the latter is composed of three modules, the compressor, the event-improved module, and the up-sampling. Their illustration is demonstrated in Figure 1 and their specific architecture is shown in Table 2. For clarity, the third column in the table illustrates the output feature map dimensions of each module when the input batch size is 1 and the length comprises 28,800 sequences. The meanings of the modules in Table 2 are as follows: Conv$(1, 256, 20)$ represents a 1D convolutional block with a kernel length of 20, 1 input channel, and 256 output channels. Convld Block$(5, \rightarrow)$ and Convld Block$(5, \leftarrow)$ denote the operation of the compressor and the up-sampling in Section 3.2.1, with 5 indicating the inclusion of 5 SCB featuring residuals. Convld$(256, 1, 1)$ signifies a convolutional block with 256 input channels and 1 output channel. Dense$(450, 128)$ denotes a fully connected layer with 450 input nodes and 128 output nodes. ConvT$(2 \times 256, 1, 20)$ represents two transposed convolution layers, each with a length of 20, 256 input channels and 1 output channel.

Table 2. Network structure and parameters.

Number	Layer Description	Output Size
	Encoder	
1	Conv$(1, 256, 20)$	(256, 3600)
	Compressor	
1	Convld Block$(5, \rightarrow)$	$(256, 1800)$
2	Convld Block$(5, \rightarrow)$	(256, 900)
3	Convld Block$(5, \rightarrow)$	(256, 450)
	DV	
1	Convld$(256, 1, 1)$+GLN+ReLU	(1, 450)
2	Dense$(450, 128)$+sigmoid	(1, 128)
3	Dense$(128, 6)$+softmax	(1, 6)
	SV	
1	Convld$(256, 1, 1)$+GLN+ReLU	(1, 450)
2	Dense$(450, 128)$+sigmoid	(1, 128)
3	Dense$(128, 6)$+softmax	(1, 6)
4	Dense$(6, 128)$+sigmoid	(1, 128)
5	Dense$(128, 450)$+sigmoid	(1, 450)
6	Convld$(256, 1, 1)$+GLN+ReLU	(256, 450)
	Up-sample	
1	Convld Block$(5, \leftarrow)$	$(256, 900)$
2	Convld Block$(5, \leftarrow)$	(256, 1800)
3	Convld Block$(5, \leftarrow)$	(256, 3600)
	Decoder	
1	Convld $(256, 2 \times 256, 1)$+ReLU	$(2 \times 256, 3600)$
2	Convld$(2 \times 256, 1, 20)$+sigmoid	(2, 28,800)

4. Experimental Procedures

4.1. Data Collection and Dataset Setup

This study conducted an experimental validation of the model's separation and recognition capabilities using a custom dataset, which encompasses both hydrophone noise and accelerometer-derived noise. For noise data collection, the experimental site is located in Danjiangkou Reservoir, Henan Province. The specific coordinates of the experimental location are 32°46.74 N, 111°33.45 E; the lake depth is 32 m; the lake surface is calm, the air temperature is 27 °C, and no other vessels passed by during data collection. As shown in Figure 4, we used a 50 kg class UUV as the object of self-noise data collection. The UUV dives to a depth of 10 m and travels at a speed of 2–4 knots. The data acquisition (DAQ) system utilized in this experiment comprises two accelerometer sensors and two underwater acoustic transducer arrays. The parameters of the data acquisition system are detailed in Table 3. The sensor arrangement is illustrated in Figures 5–7. As shown in Figures 5 and 6, two accelerometer sensors are positioned in the head and center of the UUV, respectively. In Figures 5 and 7, it can bee see that an 8-elements sensor underwater acoustic transducer array is located on both left and right flanks of the UUV, and these arrays have already been vulcanized onto the UUV's sides. In this paper, We focus on analyzing our algorithm using just use one of the underwater acoustic transducer arrays. The data acquisition process was segmented into three distinct phases: activation, uplift, and float, as shown in Figure 8. The 'activation' phase corresponds to the UUV's startup sequence, 'uplift' denotes the ascent of the UUV to the surface without propulsion, and 'float' describes the UUV's stationary presence on the water's surface. Data categorization was conducted on two levels. The primary level differentiates between sensor types: vibration noise originating from accelerometers and acoustic noise captured by the hydrophone array. The secondary level further distinguishes between the types of noise based on the experimental phases. Variations across different channels within the same sensor type were not considered in this categorization.

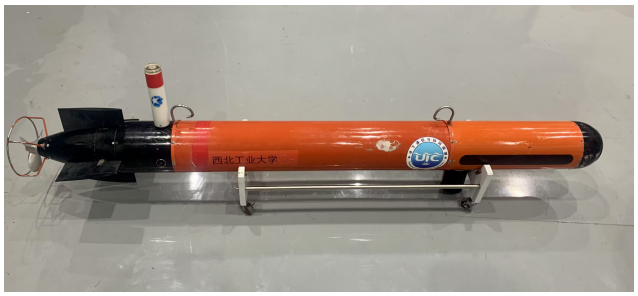


Figure 4. UUV model diagram.

Table 3. Data acquisition system introduction.

	Number	Model	Sensitivity	Usable Frequency Range	Sampling Frequency	Location
Accelerometer	2	Three-axis sensor	100 mV/g	0.4–15 kHz	48 kHz	Head and center of the UUV
Underwater acoustic transducer array	2	Eight-element sensors in each array	−200 dB	5–20 kHz	48 kHz	UUV flank array

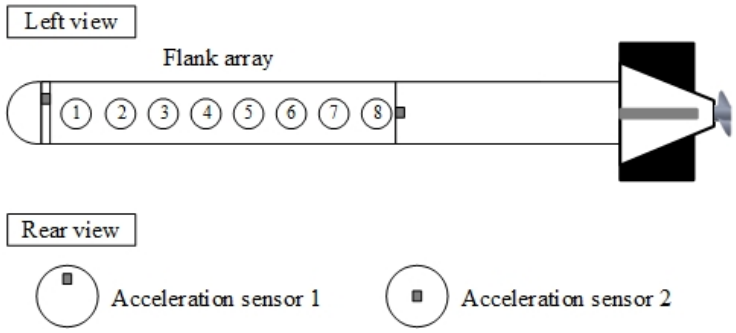


Figure 5. The sensors' arrangement.

Figure 6. Accelerometer sensor layout diagram (in the red box).

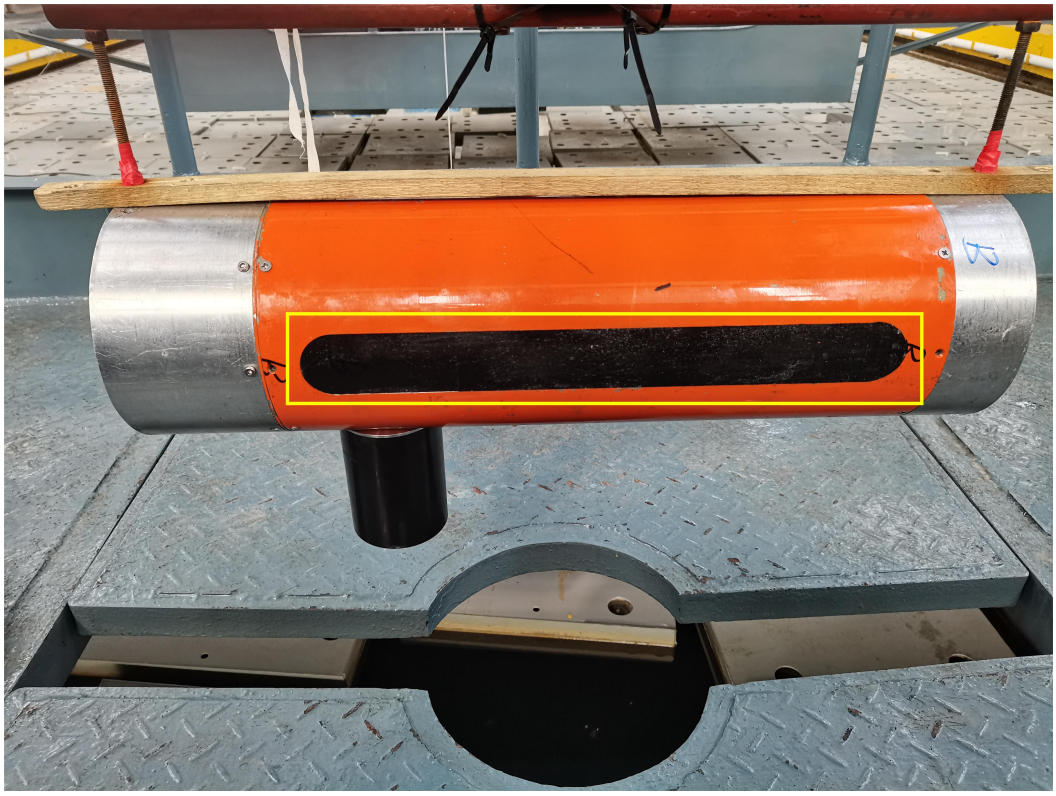


Figure 7. Underwater acoustic transducer array (in the yellow box).

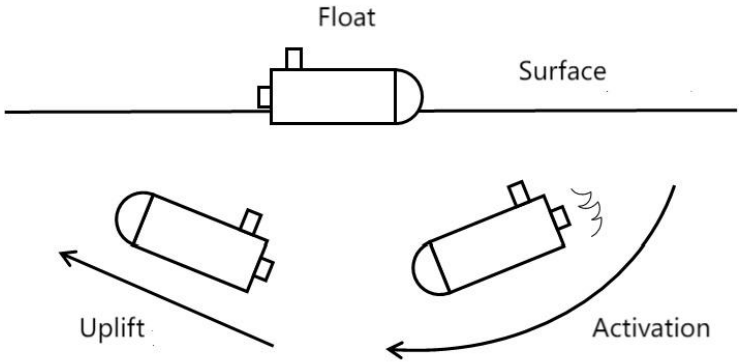


Figure 8. The data acquisition process.

Additionally, the impact of environmental noise during the UUV acquisition process is evaluated. The spectrum of the environmental noise is analyzed. As illustrated in Figure 9, the spectrum of the environmental noise is relatively low throughout the entire passband, averaging 50 dB lower than the self-noise. When the UUV is operational, all environmental noise will be masked by the self-noise, indicating that the background noise is not a primary factor influencing the UUV's performance and can be disregarded. The two-tiered noise categorization schema is summarized in Table 4.

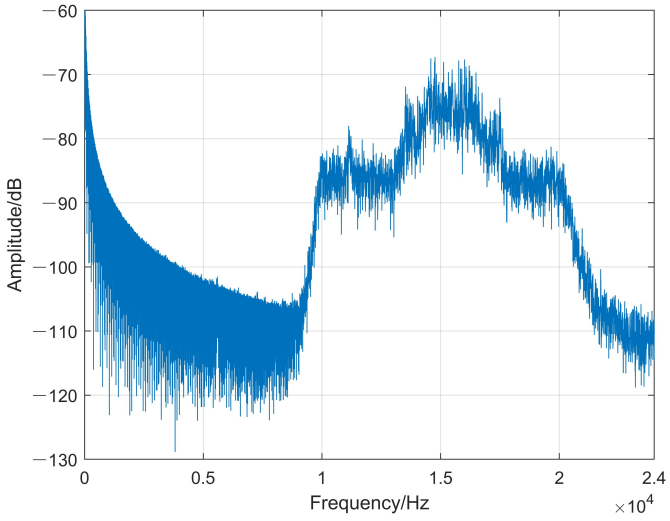


Figure 9. Environmental noise spectrum analysis result.

Table 4. The classes of sample in dataset and the size of data.

Source	Period	Number
Accelerometer	Activation	2508
	Uplift	891
	Float	2224
Hydrophone	Activation	1952
	Uplift	777
	Float	1822

Each single sample in dataset spans a duration of 0.6 s.

For both training and testing purposes, all data were sampled at a frequency of 48 kHz and divided into sequential segments, each lasting 0.6 s. To generate the mixtures, segments from various sources and periods were randomly selected and combined, with the signal-

to-noise ratios (SNR) varying from −5 to 5 dB. The distribution of the dataset for training, validation, and testing followed a 7:2:1 ratio, respectively.

4.2. Data Augmentation

In [48], a novel data augmentation technique known as dynamic mixing is employed, designed to produce a theoretically unlimited array of mixtures by randomly selecting individual samples and applying a variable SNR gain. We have adapted this dynamic mixing approach to suit our specific experimental context. For instance, during the mixing process, we select segments from two distinct periods and sources to create a mixture pair (e.g., noise from an accelerometer during the 'uplift' phase paired with noise from a hydrophone during the 'float' phase). A gain value is then randomly chosen within the range of −5 to 5 dB. By applying this gain to the selected mixture pair, we generate a single augmented mixture sample. Our experimental results are comprehensively reported, incorporating this dynamic mixing method for data augmentation.

4.3. Experiment Configurations

To accommodate the boundless possibilities afforded by dynamic mixing, we define a single epoch as comprising 4000 iterations, a decision dictated by practical considerations. The training regimen extends over 40 epochs, with each data segment processed in 0.6 s intervals. We initiate the training with a learning rate set at 1×10^{-3}. Should there be no observable enhancement in separation performance on the validation dataset for three successive epochs, the learning rate is subsequently reduced by half. The training protocol is designed to terminate prematurely if no improvement in validation performance is detected after eight continuous epochs. For optimization, we utilize the Adam algorithm, adhering to the default parameters specified within the PyTorch 2.0.1.

4.4. Evaluation Metrics

The separation method proposed in this paper belongs to a multi-task method with two outputs, one is the separated time-domain waveform and the other is the energy ratios of each single source component in mixed noise. We allocate distinct assessment methodologies for each of the defined goals. For the separated waveform, we evaluate the restoration of the waveform using both time-domain and frequency-domain metrics. Specifically, we employ the improvement of SISNR, as outlined in Equation (14), as our time-domain metric. Additionally, we utilize the line spectrum shown in Equation (16) as our frequency-domain metric, comparing the original and separated signals.

$$\text{Ratio}_l = \frac{N_l - \hat{N}_l}{N_l}, \tag{16}$$

where N_l represents the number of line spectra in the original noise, and $\hat{N}_l$ denotes the number of line spectra in the separated noise that coincide with the positions of the original noise line spectra.

To assess the recognition of the output category and energy rate of the task, we employ the mean-squared error (MSE) for comparison. The mathematical formula of this metric is presented below:

$$\text{Ratio}_e = \frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} (l_{nc} - p_{nc})^2, \tag{17}$$

where N is the number of samples, C is the number of mixed self-noise categories, and l_{nc} and p_{nc} are the energy ratio labels and network estimates for the separated category c in the sample n, respectively.

5. Result

In this section, we delineate the outcomes of our experimental evaluation, contrasting the efficacy of our proposed multi-task learning framework against alternative methodologies. Additionally, we explore the influence of various model components and parameter configurations on the overall performance.

5.1. Impact of Loss Function Weights on Performance

In this experimental series, we investigate the model's convergence behavior and its capability to accurately perform separation and recognition tasks under varying loss function weights. The multi-task learning framework is subjected to a series of tests with different emphasis on the loss components, aiming to ascertain the most effective weight distribution for optimal model training. Utilizing a grid search approach, we assess the model's performance across five distinct weight scenarios, specifically $\beta = 0.1$, $\beta = 0.3$, $\beta = 1$, $\beta = 3$ and $\beta = 10$. The evaluation employs two key metrics: the SISNR for assessing separation effectiveness and cross-entropy loss for evaluating recognition accuracy. These metrics serve to gauge the impact of varying loss function weights on the Multi-Task Learning Network (MLN)'s training efficacy.

Table 5 shows the performance scores of the model trained with MLN under five different weight values. Each row represents the loss function values for the separation and recognition tasks, while each column represents the loss function values for a single task under the five weight values. The highest performance in each task is highlighted in bold.

From the table, it can be seen that the changes in weight affect both tasks. As the weight increases from 0.1, the network's performance in both separation and recognition tasks improves. When the weight is 3, the network's separation performance reaches a peak of 13.24 dB, and the loss value for the recognition task is 0.29. Further increasing the weight to 10 causes the network's separation performance to decrease to 12.95 dB, while the recognition task's performance slightly improves to 0.28. This indicates that the network achieves a balance between the separation and recognition tasks when the weight is 3, and changing the weight either up or down reduces the overall performance of the network. Therefore, in subsequent experiments, the weight of the loss function for MLN is set to 3.

Table 5. Performance under different weights.

β	SISNR (dB)	Cross-Entropy
0.1	12.10	0.40
0.3	12.65	0.39
1	12.84	0.36
3	**13.24**	0.29
10	12.95	**0.28**

This set of experiments was conducted to study the performance of the MLN network under different configurations. In this experiment, while maintaining the rest of the network structure unchanged, new networks are formed by removing small modules from different positions each time. Using the method of controlling variables, the impact of removing each small module on the separation and estimation performance of MLN is tested. The evaluation considers three distinct configurations of the MLN, each characterized by the presence or absence of an Instance Normalization (IN) module and its impact on network functionality. Specifically, three modes are considered.

(1) MLN-N: This configuration lacks an IN module between the compression and decompression stages. As a result, the network operates strictly as a conventional separation network, yielding only a separation outcome without facilitating recognition tasks.

(2) MLN-DV: In the Diminished Version (DV) configuration, an IN module is integrated between the compression and decompression stages, transforming the network into

a multi-task learning system. This setup enables the network to produce outputs for both separation and recognition tasks. However, it does not incorporate the regularization of the latent variable space.

(3) MLN-SV: The SV configuration includes an IN module situated between the compression and decompression stages. This arrangement not only allows the network to deliver outputs for separation and recognition tasks but also implements the regularization of the latent variable space. This is achieved by utilizing task-specific labels to refine the separation process, thereby enhancing overall performance.

Table 6 shows the performance of MLN in separation and recognition tasks under different configurations with real UUV self-noise. The best performance for each metric is highlighted in bold. Without the IN module, the network only performs the separation task, outputting the time-domain waveform. The mean-squared error accumulation for the recognition task is calculated based on the energy proportion of the separated waveform. As shown in Table 6, MLN-N has the fewest parameters and the weakest performance in the three metrics, with SISNR improvements (SISNRi), line spectral error rate, and mean-squared error accumulation being 11.25 dB, 31.1%, and 0.22, respectively. MLN-DV, which adds recognition output on top of MLN-N, has the smallest mean-squared error accumulation of energy proportion among the three modes, at 0.07. Its performance in the separation task is close to that of MLN-N, with SISNRi at 11.36 dB and line spectral error rate at 28.2%. MLN-SV has the best overall performance among the three modes. Its performance in the separation task significantly improves compared to MLN-N and MLN-DV, with SISNRi increasing by 1.5 dB and line spectral error rate decreasing by 9.2%. In the recognition task, it is slightly inferior to MLN-DV, with a mean-squared error accumulation of 0.10. Overall, in the real noise set, MLN-N, MLN-DV, and MLN-SV all show a significant decrease in line spectrum estimation and prove that the IN module is effective in the real noise set.

Table 6. Performance under different modes.

Index	Separation		Recognition	Parameter (M)
	SISNR (dB)	**Ratio$_l$**	**Ratio$_e$**	
MLN-N	11.25	31.1%	0.22	2.73
MLN-DV	11.36	28.2%	**0.07**	2.91
MLN-SV	**12.86**	**19.0%**	0.10	3.05

Figure 10 presents a scatter plot of the input SISNR versus the corresponding output SISNR for each sample in the test set using MLN-SV. In the scatter plot, the input SISNR is on the horizontal axis, and the output SISNR is on the vertical axis. Each point in the plot represents the input and output SISNR for a sample during an inference. The color gradient from red to blue indicates the density of the samples, with redder points representing higher sample density and bluer points representing lower sample density. The input SISNR in Figure 10 ranges from −10 dB to 25 dB. This does not contradict the mixing SISNR sampled from the range of [−5 dB, 5 dB] because there is a correlation between different types of self-noise, and the actual input SISNR might differ from the sampled SISNR. The input SISNR here refers to the actual input SISNR.

From Figure 10, it can be observed that a large number of sample points are concentrated in the range where the input SISNR is between [−5 dB, 2.5 dB] and the output SISNR is between [7.5 dB, 15 dB].

Some sample points with high input SISNR also have higher output SISNR, while a very small number of outliers have lower output SISNR than input SISNR. This demonstrates that MLN-SV maintains stable separation output capabilities even with varying input SISNR.

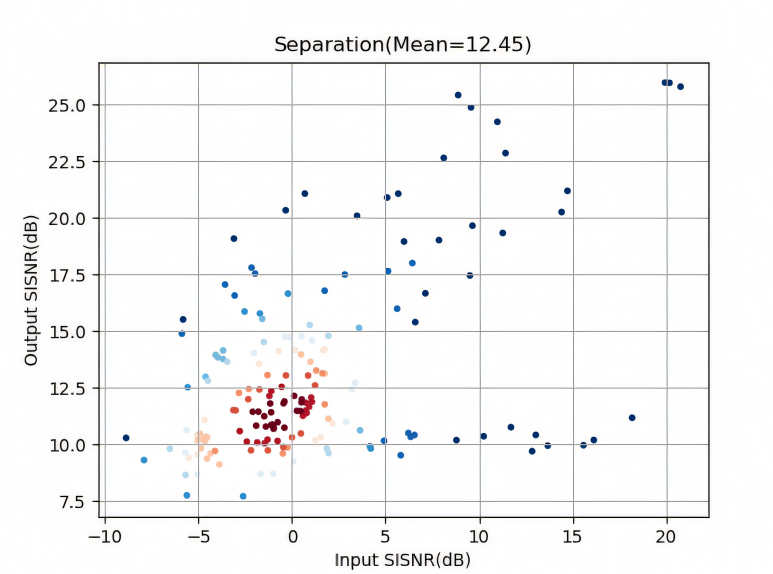


Figure 10. Scatter plot of MLN-SV input SISNR and output SISNR.

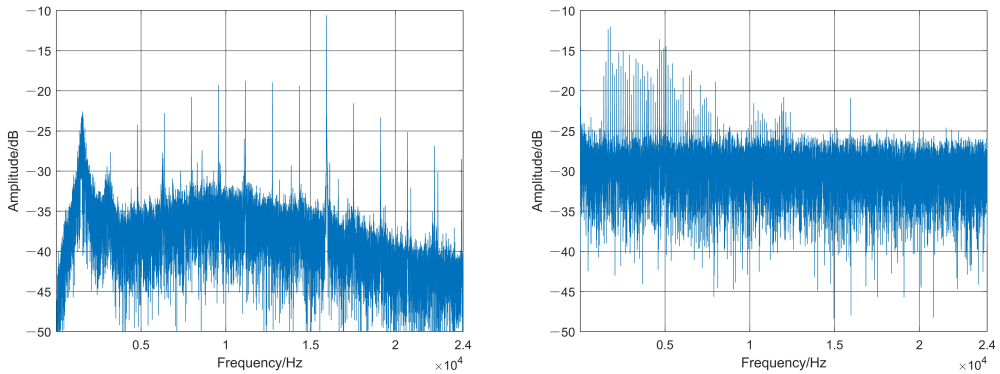
Table 7 compares the performance of MLN-SV and Conv-Tasnet. Although MLN-SV has 0.69 MB more parameters than Conv-Tasnet, it surpasses Conv-Tasnet in terms of SISNRi, line spectral error rate, and mean-squared error accumulation by 0.99 dB, 1.5%, and 0.05, respectively. This indicates that the overall performance of MLN-SV is superior to Conv-Tasnet, demonstrating stronger effectiveness and stability.

Table 7. The comparison of the performance of MLN with other models.

Index	Separation		Recognition	Parameter (M)
	SISNR (dB)	Ratio_l	Ratio_e	
MLN-SV	12.86	19.0%	0.10	3.05
Conv-Tasnet	11.87	20.5%	0.15	2.36

5.2. Separation Result

Figure 11 depicts the original activation noise and the original uplift noise, respectively. Figure 12 presents the mixed noise with an energy mixing ratio of −3 dB. From Figures 11 and 12, it can be seen that after mixing the self-noise, the frequency-domain components of the uplift noise are clearly visible, while the frequency-domain components of the activation noise nearly disappear.



(a) (b)

Figure 11. The two kinds of original self noise: (**a**) The original activation noise; (**b**) The original uplift noise.

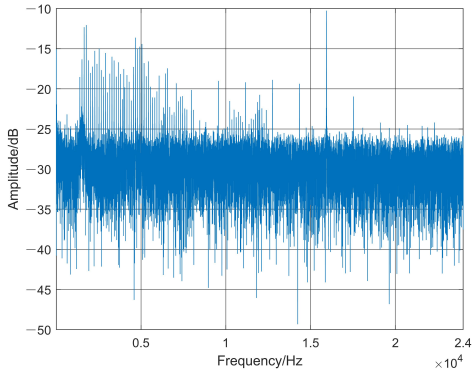


Figure 12. The mixed noise.

The mixed noise in Figure 12 was separated using MLN-DV, MLN-SV, and Conv-Tasnet, respectively. The separation results are shown in Figures 13–15. In each figure, the left subfigure shows the frequency-domain separated result for activation noise, and the right subfigure gives the frequency-domain separated result for uplift noise. Clearly, both MLN-SV and Conv-Tasnet can effectively recover the line spectrum components of the two types of noise, with MLN-SV demonstrating better recovery in the continuous spectrum compared to Conv-Tasnet. The separation results of MLN-DV are noticeably weaker than those of the other two network structures. These conclusions are consistent with the results in Tables 6 and 7.

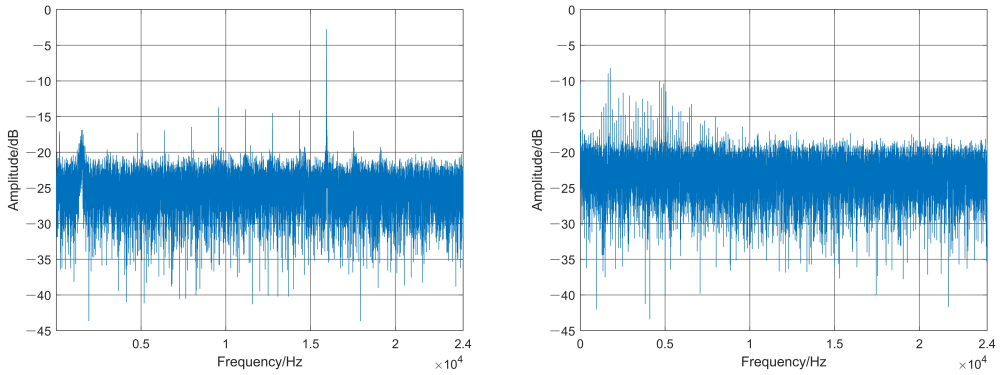


Figure 13. The separation results for MLN-DV: (**a**) The activation noise; (**b**) The uplift noise.

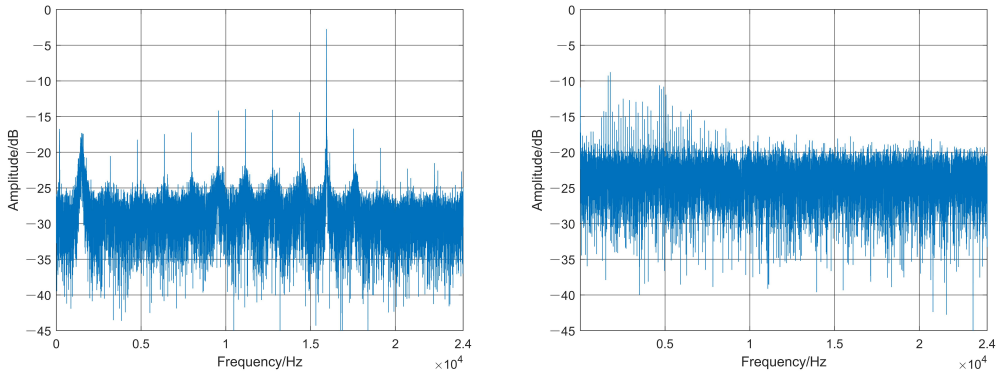


Figure 14. The separation results for MLN-SV: (**a**) The activation noise; (**b**) The uplift noise.

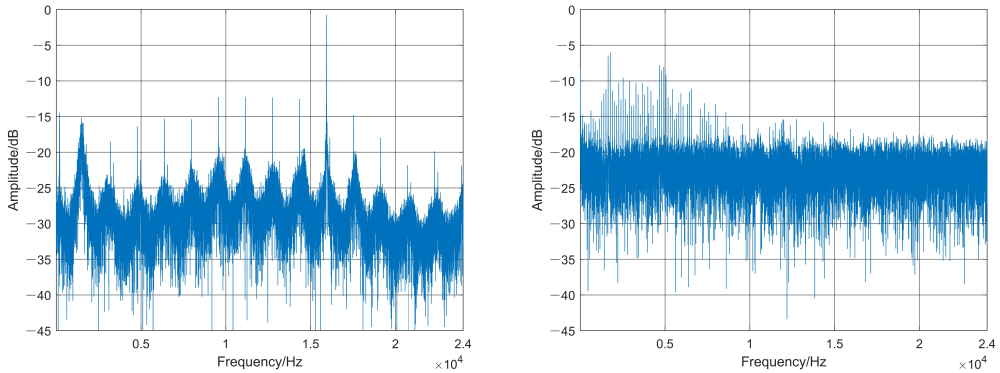


Figure 15. The separation results for Conv-Tasnet: (**a**) The activation noise; (**b**) The uplift noise.

In summary, the IN module can significantly enhance the network's separation and recognition performance. MLN-SV also demonstrates the best capability in performing time-domain waveform separation and energy ratio estimation tasks for UUV self-noise separation.

6. Conclusions

This paper explores deep learning-based algorithms for Unmanned Underwater Vehicle (UUV) self-noise separation, introducing a novel Multi-Task Learning Network (MLN) specifically designed for this task. Central to this approach is an encoder–decoder structure enhanced by an under-masking technique and a mask estimation module. By integrating a recognition task into the separation framework, the network leverages self-noise category information to refine the latent variable space, thus enhancing separation performance through informed recognition. The efficacy of the proposed MLN model is rigorously evaluated through comprehensive experimentation. Utilizing self-noise data from lake trials, the study investigates the impact of varying loss function weights on the MLN model, identifying the optimal configuration for maximum performance. Ablation tests further elucidate the contributions of different model components, affirming the algorithm's robustness and effectiveness. Moreover, comparative analyses with the Conv-Tasnet model highlight the proposed algorithm's superior performance across both Lake-Up test datasets, validating its potential for practical UUV noise management applications.

Author Contributions: Conceptualization, W.S. and D.C.; methodology, W.S. and D.C.; software, D.C. and S.L.; validation, W.S., D.C. and S.L.; formal analysis, D.C. and S.L.; investigation, W.S. and D.C.; resources, W.S. and F.T.; data curation, W.S. and F.T.; writing—original draft, W.S. and D.C.; writing—review and editing, W.S., D.C. and L.J.; visualization, W.S., D.C. and S.L.; supervision, W.S., F.T. and L.J.; project administration, W.S., F.T. and L.J.; funding acquisition, W.S. and L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under Grant No. 62371393 and 62471397, the Stable Supporting Fund of Acoustic Science and Technology Laboratory under Grant No. TCKYS2024604SST3010, and the Fundamental Research Funds for the Central Universities (23GH02027).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper are available by contacting the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their careful reading and valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UUV	Unmanned Underwater Vehicle
SINR	Signal-to-Interference-plus-Noise Ratio
DNNs	Deep Neural Networks
MLN	Multi-Task Learning Network

References

1. Cancilliere, F.M. Advanced UUV Technology. In Proceedings of the IEEE OCEANS'94, Brest, France, 13–16 September 1994; Volume 1, pp. I/147–I/151.
2. Wibisono, A.; Piran, M.J.; Song, H.K.; Lee, B.M. A Survey on Unmanned Underwater Vehicles: Challenges, Enabling Technologies, and Future Research Directions. *Sensors* **2023**, *17*, 7321. [CrossRef]
3. Liu, F.; Ma, Z.; Mu, B.; Duan, C.; Chen, R.; Qin, Y.; Pu, H.; Luo, J. Review on Fault-tolerant Control of Unmanned Underwater Vehicles. *Ocean Eng.* **2023**, *285*, 115471 [CrossRef]
4. Li, J.; Zhang, G.; Jiang, C.; Zhang, W. Review A survey of maritime unmanned search system: Theory, Applications and Future Directions. *Ocean Eng.* **2023**, *285*, 115359. [CrossRef]
5. Holmes, J.D.; Carey, W.M.; Lynch, J.F. An Overview of Unmanned Underwater Vehicle Noise in the Low to Mid Frequency Bands. *J. Acoust. Soc. Am.* **2010**, *127*, 1812. [CrossRef]
6. Kumar, P.; Ali, M.; Nathwani, K. Self-Noise Cancellation in Underwater Acoustics using Deep Neural Network Frameworks. In Proceedings of the OCEANS 2023, Limerick, Ireland, 5–8 June 2023.
7. Raanan, B.-Y.; Bellingham, J.; Zhang, Y.; Kemp, M.; Kieft, B.; Singh, H.; Girdhar, Y. Detection of Unanticipated Faults for Autonomous Underwater Vehicles Using Online Topic Models. *J. Field Robot.* **2018**, *5*, 705–716. [CrossRef]
8. Liu, F.; Tang, H.; Qin, Y.; Duan, C.; Luo, J.; Pu, H. Review on Fault Diagnosis of Unmanned Underwater Vehicles. *Ocean Eng.* **2022**, *243*, 110290. [CrossRef]
9. Zhou, A.; Zhang, W.; Li, X.; Xu, G.; Zhang, B.; Ma, Y.; Song, J. A Novel Noise-Aware Deep Learning Model for Underwater Acoustic Denoising. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4202813. [CrossRef]
10. Liu, J.; Li, Q.; Shang, D. The Investigation on Measuring Source Level of Unmanned Underwater Vehicles. In Proceedings of the 2016 IEEE/OES China Ocean Acoustics (COA), Harbin, China, 9–11 January 2016.
11. Zimmerman, R.; D'Spain, G.L.; Chadwell, C.D. Decreasing the Radiated Acoustic and Vibration Noise of a Mid-Size AUV. *IEEE J. Ocean. Eng.* **2005**, *1*, 179–187. [CrossRef]
12. Soni, S.; Yadav, R.N.; Gupta, L. State-of-the-Art Analysis of Deep Learning-Based Monaural Speech Source Separation Techniques.*IEEE Access* **2023**, *11*, 4242–4269. [CrossRef]
13. Gu, J.J.; Yao, D.D.; Li, J.F.; Yan, Y.H. A Novel Semi-Blind Source Separation Framework towards Maximum Signal-to-Interference Ratio. *Signal Process.* **2024**, *217*, 109359. [CrossRef]
14. Drude, L.; Haeb-Umbach, R. Integration of Neural Networks and Probabilistic Spatial Models for Acoustic Blind Source Separation. *IEEE J. Sel. Top. Signal Process.* **2019**, *4*, 815–826. [CrossRef]
15. Chen, J.; Liu, C.; Xie, J.W.; An, J.; Huang, N. Time-Frequency Mask-Aware Bidirectional LSTM: A Deep Learning Approach for Underwater Acoustic Signal Separation. *Sensors* **2022**, *15*, 5598. [CrossRef] [PubMed]
16. Ansari, S.; Alatrany, A.S.; Alnajjar, K.A.; Khater, T.; Mahmoud, S.; Al-Jumeily, D.; Hussain, A.J. A Survey of Artificial Intelligence Approaches in Blind Source Separation. *Neurocomputing* **2023**, *561*, 126895. [CrossRef]
17. Yu, Y.S.; Qiu, X.Y.; Hu, F.C.; He, R.H.; Zhang, L.K. An End-to-End Speech Separation Method Based on Features of Two Domains. *J. Vib. Eng. Technol.* **2024**, *12*, 7325–7334. [CrossRef]
18. Song, R.P.; Feng, X.; Wang, J.F.; Sun, H.X.; Zhou, M.Z.; Esmaiel, H. Underwater Acoustic Nonlinear Blind Ship Noise Separation Using Recurrent Attention Neural Networks. *Remote Sens.* **2024**, *4*, 653. [CrossRef]
19. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *8*, 1256–1266. [CrossRef] [PubMed]
20. Chen, J.; Mao, Q.; Liu, D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020.
21. Hu, X.; Li, K.; Zhang, W.; Luo, Y.; Lemercier, J.-M.; Gerkmann, T. Speech Separation Using an Asynchronous Fully Recurrent Convolutional Neural Network. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021.
22. Défossez, A.; Usunier, N.; Bottou, L.; Bach, F. Music Source Separation in the Waveform Domain. *arXiv* **2021**, arXiv:1911.13254.

23. Hung, Y.-N.; Lerch, A. Multitask Learning for Instrument Activation Aware Music Source Separation. *arXiv* **2020**, arXiv:2008.00616.
24. Lee, K.J.; Lee, B. End-to-End Deep Learning Architecture for Separating Maternal and Fetal ECGs Using W-Net. *IEEE Access* **2022**, *10*, 39782–39788. [CrossRef]
25. Tzinis, E.; Wisdom, S.; Hershey, J.R.; Jansen, A.; Ellis, D.P.W. Improving Universal Sound Separation Using Sound Classification. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020.
26. Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K.; Roux, J.L.; Hershey, J.R. Universal Sound Separation. In Proceeding of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019.
27. Wisdom, S.; Tzinis, E.; Erdogan, H.; Weiss, R.J.; Wilson, K.; Hershey, J.R. Unsupervised Sound Separation Using Mixture Invariant Training. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual, 6–12 December 2020.
28. Singh, A.; Ogunfunmi, T. An Overview of Variational Autoencoders for Source Separation, Finance, and Bio-Signal Applications. *Entropy* **2022**, *1*, 55. [CrossRef]
29. Wang, D.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *10*, 1702–1726. [CrossRef] [PubMed]
30. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2020**, arXiv:1312.6114. [CrossRef]
31. Li, L.; Kameoka, H.; Makino, S. FastMvAE2: On Improving and Accelerating the Fast Variational Autoencoder-Based Source Separation Algorithm for Determined Mixtures. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 96–110. [CrossRef]
32. Casti, P.; Cardarelli, S.; Comes, M.C.; D'Orazio, M.; Filippi, J.; Antonelli, G.; Mencattini, A.; Di Natale, C.; Martinelli, E. S3-VAE: A Novel Supervised-Source-Separation Variational AutoEncoder Algorithm to Discriminate Tumor Cell Lines in Time-Lapse Microscopy Images. *Expert Syst. Appl.* **2023**, *232*, 120861.1–120861.15. [CrossRef]
33. Pal, M.; Roy, R.; Basu, J.; Bepari, M.S. Blind Source Separation: A Review and Analysis. In Proceedings of the 16th International Oriental COCOSDA Conference, Gurgaon, India, 25–28 November 2013.
34. Heurtebise, A.; Ablin, P.; Gramfort, A. Multiview Independent Component Analysis with Delays. In Proceedings of the 33rd IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2023), Rome, Italy, 17–20 September 2023.
35. Yoshii, K.; Tomioka, R.; Mochihashi, D.; Goto, M. Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction. In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013), Curitiba, Brazil, 4–8 November 2013; pp. 369–374.
36. Parekh, J.; Parekh, S.; Mozharovskyi, P.; Richard, G.; d'Alché-Buc, F. Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1392–1405. [CrossRef]
37. Do, H.D.; Tran, S.T.; Chau, D.T. Speech Source Separation Using Variational Autoencoder and Bandpass Filter. *IEEE Access* **2020**, *8*, 156219–156231. [CrossRef]
38. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P. Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *12*, 2136–2147. [CrossRef]
39. Hershey, J.R.; Chen, Z.; Le Roux, J.; Watanabe, S. Deep Clustering: Discriminative Embeddings for Segmentation and Separation. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 31–35.
40. Yu, D.; Kolbæk, M.; Tan, Z.H.; Jensen, J. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 241–245.
41. Kolbæk, M.; Yu, D.; Tan, Z.H.; Jensen, J. Multitalker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *10*, 1901–1913. [CrossRef]
42. Pishdadian, F.; Wichern, G.; Le Roux, J. Finding Strength in Weakness: Learning to Separate Sounds with Weak Supervision. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2386–2399. [CrossRef]
43. Seetharaman, P.; Wichern, G.; Venkataramani, S.; Roux, J.L. Class-Conditional Embeddings for Music Source Separation. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 301–305.
44. Karamatlı, E.; Cemgil, A.T.; Kırbız, S. Weak Label Supervision for Monaural Source Separation Using Non-Negative Denoising Variational Autoencoders. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019.
45. Grais, E.M.; Plumbley, M.D. Single Channel Audio Source Separation Using Convolutional Denoising Autoencoders. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP) , Montreal, QC, Canada, 14–16 November 2017; pp. 1265–1269.
46. Bofill, P.; Zibulevsky, M. Underdetermined Blind Source Separation Using Sparse Representations. *Signal Process.* **2001**, *11*, 2353–2362. [CrossRef]

47. Li, L.; Kameoka, H.; Makino, S. Fast MVAE: Joint Separation and Classification of Mixed Sources Based on Multichannel Variational Autoencoder with Auxiliary Classifier. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 546–550.
48. Li, C.; Luo, Y.; Han, C.; Li, J.; Yoshioka, T.; Zhou, T.; Delcroix, M.; Kinoshita, K.; Boeddeker, C.; Qian, Y.; et al. Dual-path RNN for Long Recording Speech Separation. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 865–872.