SJTU-AUDIOCC SYSTEM FOR DCASE 2025 CHALLENGE TASK 4: SPATIAL SEMANTIC SEGMENTATION OF SOUND SCENES

Technical Report

*Xin Zhou, Hongyu Wang, Chenda Li, Bing Han, Xinhu Zheng, Yanmin Qian**

Auditory Cognition and Computational Acoustics Lab
MoE Key Lab of Artificial Intelligence, AI Institute
School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

The present report introduces four systems developed by the AudioCC Lab at Shanghai Jiao Tong University for DCASE 2025 Task 4. The task at hand is to detect target sound events and separate corresponding signals from multi-channel mixture. It was found that the effective detection of sound events and extraction of the corresponding signals was challenging under conditions where mixture consists of multiple target sound events, non-target sound events, and non-directional background noise. In order to address this challenge, we propose four systems. The first system represents an enhancement to the baseline system, the second is a multi-stage iterative system that is both novel and promising, the third is a lightweight model based on Encoder-Decoder Attractor (EDA) module, and the fourth integrates multiple audio tagging models to achieve optimal performance. These four systems cover high performance, low overhead, and promising frameworks, providing a reference for future research on this task.

***Index Terms*—** audio tagging, label-queried sound separation, iterative method, encoder-decoder attractor, model ensemble

1. INTRODUCTION

The DCASE 2025 Task 4 [1, 2], titled "Spatial Semantic Segmentation of Sound Scenes", is centered on detecting multiple target sound events and extracting corresponding signals from multi-channel mixture.

In this challenge, the detection task [3] is simplified to an audio tagging problem, which places greater emphasis on identifying what is happening in the audio signal rather than the exact timing of the sound events. The diverse acoustic characteristics of sound events, coupled with the simultaneous occurrence of multiple sounds leading to overlapping events, pose significant challenges and degrade detection performance. To address these challenges, supervised learning models such as CED [4], PANN [5], and PaSST [6] leverage large-scale datasets with tag annotations (e.g. AudioSet [7]) for pretraining to perform classification tasks. However, these models are unable to utilize the vast amounts of unlabeled data. On the other hand, models employing unsupervised or semi-supervised learning strategies, such as BEATs [8], SSLAM [9], EAT [10], and Dasheng [11], focus on learning meaningful representations from large amounts of unlabeled data. However, these models still struggle to exhibit generalization capabilities when encountering unseen or complicated samples.

*Corresponding author.

The separation task can be transformed into a target sound extraction task based on weak labels [12, 13]. There are many well-known audio separation models currently [14, 15], but they cannot solve the label permutation problem. Therefore, some studies have attempted to utilize some prior knowledge about the target sound, which we refer to as clue (which can be a sound tag, video, audio, or text) to extract the specified audio. In this task, labels of target sound events are used for label-queried sound separation. In addition, multi-channel audio can provide additional spatial information to help separate different sound sources.

In this technical report, we emphasize that we have made the following contributions to the DCASE 2025 Task 4.

- Explores the effectiveness of supervised and unsupervised pre-trained audio tagging models in this challenge and further attempts to fine-tune the aforementioned models on different scales to enhance performance.
- Investigates the impact of model ensembling with different combinations on audio tagging performance.
- Provides a variety of effective strategies for the DCASE 2025 Task 4.

2. DATASET

For training and validation data, the DCASE 2025 Task 4 organizers provide dry source sample files (Anechoic Sound Event 1K, newly recorded by NTT + FSD50K [16] + EARS dataset [17]), RIR files (NTT recorded + FOA-MEIR [18]), non-directional background noise and interference event sound files (FOA-MEIR + FSD50K + ESC-50 [19] + DISCO [20] that are used in Semantic Hearing [21]). The input signals are designed to contain multi-channel audio mixtures with up to three simultaneous target sound events, along with optional multiple non-target sound events and non-directional background noise. Each output signal is expected to contain one isolated target sound event with a predicted label for the event class.

The audio tagging models were pre-trained on large-scale general audio source datasets such as Audioset [7] and VGGSound [22]. Audioset contains more than 2 million 10-second audio segments manually weakly labeled from YouTube, covering 632 audio event categories, including a wide range of human and animal sounds, musical instruments and music genre sounds, everyday environmental sounds, and so on. VGGSound covers 310 audio categories, including human voices, natural sounds, musical instruments, and more. It contains more than 210,000 video segments, each lasting 10 seconds and sourced from YouTube, with a total

duration of over 550 hours.

3. SYSTEM DESCRIPTION

3.1. Baseline system

The organizers have provided two baseline systems: (a) M2D [23] + ResUNet [24] and (b) M2D + ResUNetK. Baseline (a) employs an audio tagging model M2D, which takes the first channel of the multi-channel mixture as input and extracts 1–3 labels from the output probabilities. These label vectors are then fed into a ResUNet model that accepts the spectrogram of the multi-channel mixture as input. For each predicted label, ResUNet separately extracts the corresponding sound source. Baseline (b) differs from (a) in that the ResUNetK model takes the concatenation of multiple one-hot encoded label vectors as input and processes all predicted labels simultaneously.

3.2. System 1: multi-channel M2D (AT) + SepformerK (LSS)

We adopted the M2D model for audio tagging and replaced the separation model with the Sepformer [15]. The same as the baseline, our M2D model is based on a version pre-trained on AudioSet dataset [7] at 32 kHz, and we further fine-tuned it in two stages using the dataset provided for the challenge. First, we fine-tuned the classification head, and then unlike the baseline, which fine-tuned 2 blocks, we fine-tuned the entire model. In addition, instead of using only the first channel as in the baseline, we process each channel of the multi-channel mixture with the M2D model and aggregate the resulting probabilities using the power mean method. Sepformer is a time-domain separation model. It processes long sequences by segmenting them into smaller chunks and performing operations both within and between these chunks. The model adopts a dual-path structure design and utilizes transformers to learn the short- and long-range dependencies in the audio. Here, we also imitated the strategy in ResUNetK, enabling Sepformer to predict multiple signals simultaneously based on the concatenated input label vector.

3.3. System 2: Two Stage Iterative Method

Sound separation has been used as a preprocessing step to improve the results of sound event detection [25]. And in contrast, [26] and [27] show that the performance of an sound separation network can be enhanced by incorporating the embedding information extracted by a sound classifier model. Therefore, we considered whether these two findings could be combined. To enhance the detection and separation performance, we introduced a two-stage iterative tagging and separation model as shown in Fig. 1. This model includes two taggers (Tagger 1 and Tagger 2) and two separators (Separator 1 and Separator 2). For separators, we employed band-split RNN (BSRNN, a frequency-domain separation model) [14]. Compared with Sepformer, it can be trained faster and the overhead is lower. In the first stage, Tagger 1 takes the 4-channel mixture as input and outputs probabilities for 18 classes. These probabilities are aggregated across the channel dimension by power mean and then used as input to Separator 1. Separator 1 combines the feature with the aggregated probabilities from Tagger 1 to output three separated audio sources. In the second stage, Tagger 2 takes the 4-channel mixture and the three separated sources from Separator 1 (a total of 7 channels), outputs probabilities for 18 classes, and aggregates these results through power mean method. Separator 2, during training, receives ground truth labels, the 4-channel mixture, and the three sources output by Separator 1 as extra channels, outputting sources corresponding to each label; during inference, it uses the estimated labels from Tagger 2 and 7-channel signals mentioned above as input.

To accommodate the training needs of different modules, we first used the weights of a fine-tuned M2D model as the initial weights for the two taggers and froze them. And then we trained the two separators. After this stage, we fine-tuned the entire model (including the two taggers and two separators) with a smaller learning rate to optimize the overall performance of the system.

To ensure tagging accuracy, the outputs of Tagger 1 and Tagger 2, after aggregated, are compared with the ground truth labels to compute the binary cross entropy (BCE) loss. The output of Separator 1 is compared with the reference sources to calculate the permutation invariant training (PIT) source-to-distortion ratio (SDR) loss, and the output of Separator 2 is compared with the reference sources corresponding to the labels to calculate the fixed-ordered SDR loss.

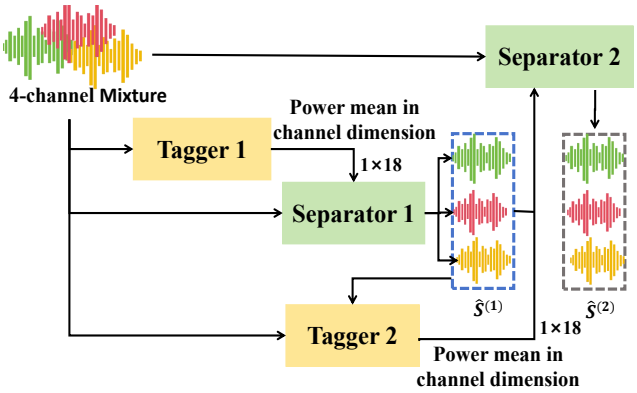

Figure 1: The structure of the two stage iterative method.

3.4. System 3: EDA module for Tagging and Separation

We employ SepEDA [28], which is built upon the Sepformer and EDA module [29]. SepEDA estimates the number of target sound sources in the mixture and the representation embeddings of each source, referred to as attractors. Each attractor is expected to represent a target sound event. The model fuses these attractors into the intermediate features respectively and finally outputs the corresponding signal for each attractor. Besides, we appended a classification head at the end of the EDA module to enable the attractor to explicitly learn semantic information about the sound event class. The separated sources are compared with the reference sources to calculate the PIT SDR loss. Then the permutation obtained from the PIT loss calculation is then used for the 18-class classification BCE loss calculation. Additionally, we calculate the BCE loss for sound source counting. It is important to note that, in contrast to other systems which utilize separate tagger and separator, System 3 employs a single model to perform both tagging and separation, which is conducive to the joint optimization of the entire system.

3.5. System 4: Ensemble Learning for Audio Tagging

In the audio tagging phase, we employed ensemble learning to further enhance the performance. Specifically, we trained multiple

Table 1: Performance Comparison of Audio Tagging Models with Finetuning When Separator is SepformerK.

Model	# Params (M)	Partial Model Finetuning		Entire Model Finetuning	
		Acc (%)	CA-SDRi (dB)	Acc (%)	CA-SDRi (dB)
BEATs	90.3	64.467	13.464	68.533	13.893
Dasheng	85.5	62.600	13.197	66.067	13.628
SSLAM	90.0	64.533	13.504	-	-
M2D	85.5	63.133	13.259	**70.467**	**14.023**
CED	85.3	64.600	13.479	66.733	13.760

models, including BEATs [8], Dasheng [11], SSLAM [9], M2D [23], and CED [4], to address the audio tagging task. Among them, CED is a supervised pre-trained model, while the other models are self-supervised pre-trained models. Each of these models was fine-tuned on the challenge-provided dataset to leverage their unique strengths and capture diverse features from the audio. To aggregate the predictions from these audio tagging models, we simply average the probabilities output from these models for each of the 18 classes.

4. EXPERIMENTAL SETUP

The experiment was conducted based on the baseline code repository, and we have made some modifications to the codes to meet our requirements. We used NVIDIA GeForce RTX 4090 and NVIDIA A10 for training, with 4 GPUs for training the separation task and 2 GPUs for training the tagging task.

4.1. Setup for four systems

For System 1, we evaluated the performance of each pre-trained model on the test set after fine-tuning them at different scales. For M2D, we followed the baseline system and fine-tuned two blocks and the entire model respectively. While for the other models, we fine-tuned four blocks and the entire model respectively.

For System 2 and 3, we have roughly introduced the training process in Section 3.

For System 4, based on the models trained in the experiments of System 1, we tried different model combinations and selected the one that performed best on the test set as our System 4.

4.2. Unreliable Validation Set

We found that the sounds of the AlarmClock class in the dry sound sources of the validation set were completely different from those of the AlarmClock class in the training set. The AlarmClock sounds in the validation set don't quite conform to the common understanding of this type of sound. They are more like musical alarms. Among the 540 mixtures in the validation set, 54 of them contain sounds of the AlarmClock class, which may significantly reduce the reliability of the validation set.

During the experiment, we did find that the performance of the model on the test set of the development dataset was not closely related to the loss on the validation set. Therefore, we referred more to the loss on the training set to select the final model. Additionally, since we tried fine-tuning the entire audio tagging model during the second stage of training, which might lead to overfitting, we did not train the model to convergence on the training set.

4.3. Evaluation Metric

In this year's task, we need to simultaneously focus on both the tagging and separation performance. Therefore, we adopted the class-aware signal-to-distortion ratio improvement (CA-SDRi) [1, 2] proposed by the organizers as the final evaluation metric. Since incorrect predictions do not contribute to any improvement in the metric, the system must first tag the mixture accurately. The CA-SDRi formula is shown in Equation 1 and Equation 2:

$$\begin{aligned}\text{CA-SDRi}&\left(\{\hat{x}_1,\ldots,\hat{x}_{\hat{K}}\},\{x_1,\ldots,x_K\},\hat{C},C,y\right)\\&=\frac{1}{|C\cup\hat{C}|}\sum_{c_k\in C\cup\hat{C}}P_{c_k},\end{aligned}\tag{1}$$

where C represents the set of ground truth sound events, $\hat{C}$ denotes the set of predicted sound events, and y is the audio mixture. The terms $\{\hat{x}_1,\ldots,\hat{x}_{\hat{K}}\}$ refer to the audio signals of the predicted individual sound events, while $\{x_1,\ldots,x_K\}$ are the audio signals of the true individual sound events. The metric component P_{c_k} is calculated as:

$$P_{c_k\in C\cup\hat{C}}=\begin{cases}\text{SDRi}(\hat{x}_k,x_k,y), & \text{if } c_k\in C\cap\hat{C}\\ \mathcal{P}^{\text{FN}}_{c_k}, & \text{if } c_k\in C \text{ and } c_k\notin\hat{C},\\ \mathcal{P}^{\text{FP}}_{c_k}, & \text{if } c_k\notin C \text{ and } c_k\in\hat{C}\end{cases}\tag{2}$$

where $\mathcal{P}^{\text{FN}}_{c_k}$ and $\mathcal{P}^{\text{FP}}_{c_k}$ are the penalty values for false negative (FN) and false positive (FP) cases, respectively. Here they are both set to 0.

5. RESULTS AND ANALYSIS

This section comprehensively investigates the performance of audio tagging models and separation models from multiple perspectives. Finally, we present the performance results of our four systems.

5.1. Audio Tagging Models with Finetuning

As shown in Table 1, we attached a classification head to the end of the pre-trained models and investigated the effects of different degrees of fine-tuning. We present the tagging accuracy of audio tagging models, and calculate the CA-SDRi metric when introducing the separator. It should be noted that due to time constraints and the unreliability of the validation set, the models may not have been trained to be optimal. And some experiments are even missing, such as fine-tuning the entire model of SSLAM. However, it can still be observed that fine-tuning the entire model can lead to better performance, although it may cause overfitting, which can be mitigated through various methods, such as weight decay and dropout.

Table 2: Performance Evaluation of Four Systems and Baselines for Audio Tagging and Separation. MC stands for multi-channel input.

System	Method Type	Tagging	Separation	# Params (M)	Acc (%)	CA-SDRi (dB)
Baseline (a)	Tagging + Separation	M2D (finetune 2 layers)	ResUNet	115.4	59.8	11.03
Baseline (b)	Tagging + Separation	M2D (finetune 2 layers)	ResUNetK	115.4	59.8	11.09
System 1	Tagging + Separation	MC-M2D (finetune all)	SepformerK	105.1	73.933	14.381
System 2	Two Stage Iterative	MC-M2D (finetune all)	BSRNN	204	62.733	12.400
System 3	EDA module	Linear head for attractors	SepEDA	8.9	49.533	10.468
System 4	Tagging + Separation	BEATs + MC-M2D	SepformerK	195.4	**76.267**	**14.657**

5.2. Audio Tagging with Single-Channel or Multi-Channel

In the baseline system, M2D only uses the first channel of the four-channel mixture. We have tried feeding each channel into the audio tagging model separately and aggregate the probabilities corresponding to each channel output by the model through the quadratic power mean method. The results in Table 3 show that the model using multi-channel audio performs better on the test set. This is reasonable because multiple channels represent additional information.

Table 3: M2D Model Performance: Single-Channel vs. Multi-Channel Audio Tagging.

Model	Acc (%)	CA-SDRi (dB)
Single-Channel M2D	70.467	14.023
Multi-Channel M2D	**73.933**	**14.381**

5.3. Ensemble Approaches for Robust Audio Tagging

To improve performance, we tried various combinations of models. The performance of some combinations is shown in Table 4. As can be seen, an increase in the number of ensemble models does not necessarily lead to an overall performance increase. Since we use simple averaging method to aggregate the outputs of individual models, models with poorer performance can have a negative impact on the entire system. Finally, based on the performance on the test set, we chose BEATs + multi-channel M2D as the submitted system 4. It should be noted that since time is limited and the validation set is not so reliable, the performance of each model we trained may not be optimal.

Table 4: Performance of Ensemble Approaches for Audio Tagging

Ensemble Configuration	Acc (%)	CA-SDRi (dB)
BEATs + MC-M2D	**76.267**	**14.657**
SSLAM + MC-M2D	74.867	14.467
BEATs + SSLAM + MC-M2D	75.533	14.600
BEATs + Dasheng + MC-M2D	74.933	14.496
All Models	73.667	14.419

5.4. Evaluation of Four Systems and Baseline: Tagging Accuracy and CA-SDRi Scores

Table 2 shows the number of parameters and the performance of the baseline systems and our four systems. With our improvement, the performance of System 1 has been significantly enhanced compared to the baseline, while the number of parameters is similar to that of the baseline system. System 4 introduces an additional tagging model, and further enhances the performance of the ensemble system. However, the number of parameters also increases significantly. Due to the time limit and the complexity of the architecture, System 2 was not trained to be optimal. However, System 2 introduces extra processing on the basis of the baseline system and should be promising. The tagging accuracy of System 3 is not good enough, which might be due to the model's insufficient ability to handle complicated mixtures. However, similarly, the system has not been fully trained and there is still room for further improvement in performance. Meanwhile, the number of parameters of SepEDA is very small, similar to that of Sepformer, which is a major advantage of this system.

For each system, especially System 2 and 3, there is a lot of room for further improvement. We may conduct some extra research in the future.

6. CONCLUSION

This technical report presents the audio tagging and separation models developed by the AudioCC Lab at Shanghai Jiao Tong University. In the audio tagging challenge, we focused on finetuning large-scale pre-trained general sound models, as illustrated in System 1. To further enhance the performance, we employed ensemble learning techniques, which are exemplified in System 4. In the audio separation task, we utilized advanced models such as Sepformer and BSRNN. Additionally, we explored a multi-stage iterative system, as demonstrated in System 2, and a lightweight model based on the EDA module in System 3. Our ensemble system achieved the highest score on the development dataset test set, with a CA-SDRi of 14.657, representing a significant improvement over the baseline of challenge.

7. REFERENCES

[1] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, *et al.*, "Description and discussion on dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2506.10676*, 2025.

[2] B. T. Nguyen, M. Yasuda, D. Takeuchi, D. Niizumi, Y. Ohishi, and N. Harada, "Baseline systems and evaluation metrics for spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2503.22088*, 2025.

[3] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley,

"Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[4] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.

[5] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[6] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[7] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[9] T. Alex, S. Atito, A. Mustafa, M. Awais, and P. J. Jackson, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," in *The Thirteenth International Conference on Learning Representations*, 2025.

[10] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[11] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," *arXiv preprint arXiv:2406.06992*, 2024.

[12] C. Li, Y. Qian, Z. Chen, D. Wang, T. Yoshioka, S. Liu, Y. Qian, and M. Zeng, "Target sound extraction with variable cross-modality clues," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2022.

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

[15] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[16] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[17] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," *arXiv preprint arXiv:2406.06185*, 2024.

[18] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 226–230.

[19] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[20] L. Lanzendörfer, F. Grötschla, E. Funke, and R. Wattenhofer, "Disco-10m: A large-scale music dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 54 451–54 471, 2023.

[21] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–15.

[22] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[23] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[24] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.

[25] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," 2020. [Online]. Available: https://arxiv.org/abs/2007.03932

[26] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, D. Niizumi, N. Tawara, T. Nakatani, and S. Araki, "Soundbeam meets m2d: Target sound extraction with audio foundation model," 2024. [Online]. Available: https://arxiv.org/abs/2409.12528

[27] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[28] S. R. Chetupalli and E. A. Habets, "Speech separation for an unknown number of speakers using transformers with encoder-decoder attractors." in *INTERSPEECH*, 2022, pp. 5393–5397.

[29] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.