A MULTI-STAGE SEPARATION-AND-CLASSIFICATION FRAMEWORK GUIDED BY COMPLEMENTARY ACOUSTIC-TO-SEMANTIC CLUES

Technical Report

Younghoo Kwon[1], *Junwoo Park*[1], *Han Yin*[1], *Jung-Woo Choi*[1*],

[1] School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
{k0hoo, park.junwoo, hanyin, jwoo}@kaist.ac.kr

ABSTRACT

This report describes the system proposed for the DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). Specifically, we develop a multi-stage framework in which each stage couples a separation model with a classification model. The first stage performs source separation and classification directly on the multi-channel mixture. Its outputs are then propagated to the following stage as two complementary clues that progressively refine each target estimate: (i) an enrollment clue, the separated waveform itself, serving as a low-level acoustic reference; and (ii) a class clue, the predicted label encoded as a one-hot vector. The third stage reuses the second-stage outputs under the same scheme, forming an iterative self-guided refinement process. In addition, we use a fine-grained frame-level audio embedding from an audio encoder pretrained on a large audio corpus as an additional clue to further improve the audio separation performance. On the test set, the proposed system achieves a CAPI-SDRi of 15.51 dB, a mixture accuracy of 71.09%, and a source accuracy of 78.62%; with an improvement of 7.02 dB, 10.38%p and 8.22%p compared with the challenge baseline, respectively.

***Index Terms*—** Audio Source Separation, Audio Classification, Duration-Based Augmentation, Temporal-FiLM

1. INTRODUCTION

Spatial Semantic Segmentation of Sound Scenes (S5) aims to detect and separate individual sound events from a multi-channel mixture. Given a recording of several directional sources, interfering sounds, and diffuse background noise, an S5 system must both recognize target classes and recover an isolated waveform for each detected source. S5 systems can be applied in various applications, such as auditory scene analysis.

Following DCASE 2025 Task 4 [1], the S5 task continues in DCASE 2026 Task 4 [2] with two modifications that bring the benchmark closer to real acoustic conditions. First, a single mixture may now contain multiple sources of the same class simultaneously. For example, several people may speak at once. Second, a mixture may contain zero target events, so that the system must reliably decide when no target sound is present while still being exposed to background noise and interference. Both changes substantially increase the difficulty of the task and require systems to be robust to source-count uncertainty and label ambiguity.

As a foundation for this task, we build upon a multi-stage self-guided framework initially developed for DeepASA [3] and DCASE 2025 Task 4 [4]. These baseline frameworks progressively refine separation and classification using internally generated clues. Especially, the framework proposed in [4] employs a Universal Sound Separation (USS) model, DeFT-Mamba-USS, to decompose multi-channel mixtures into distinct sources, followed by a Single-label Classification (SC) model, M2D-SC, to predict their labels. The separated waveforms and predicted classes are then fed into a Target Sound Extraction (TSE) model, DeFT-Mamba-TSE, as clues to form a self-guided, iterative refinement loop.

The initial baseline [4] faced notable inefficiencies in its classification model (M2D-SC). Specifically, fine-tuning the classifier on a relatively small dataset degraded the diverse representational capacity of the pretrained model. Furthermore, converting the separated waveforms into magnitude-only mel-spectrograms caused the loss of phase and fine-grained frequency information. To overcome these limitations, a Dual-Path Classifier (DPC) was proposed [5]. The DPC directly utilizes the intermediate object features produced from the USS model, alongside semantic features extracted from a frozen pretrained M2D model. This approach preserves fine-grained frequency information that is lost during Mel-spectrogram projection. Concurrently, an SEC module [5] integrates semantic embeddings from pretrained models (e.g., M2D or MGA-CLAP) into DeFT-Mamba-TSE to enrich the class clues. These embeddings are projected to match the dimensions of the one-hot class embedding vector and are added element-wise before being injected into the DeFT-Mamba-TSE blocks via FiLM layers.

Despite these advancements, applying the existing framework reveals two major limitations. First, the conventional SEC approach simply combines embeddings from one-hot class vectors and embeddings from a pretrained model through element-wise addition. This can dilute information contained in each embedding required for precise and rich target extraction. Furthermore, the embeddings from pretrained models have a coarse time resolution, failing to capture the fine-grained temporal dynamics. Second, we observe a classification performance degradation in percussive classes (e.g., percussion), which are frequently misclassified as silence. Our assumption is that this degradation is caused by the prevalence of extremely short, transient samples in the training dataset, which provide insufficient temporal context for reliable detection. In this report, we propose two methods to overcome these limitations:

- **Fine-grained Semantic Conditioning**: We utilize pretrained AF-Whisper [6] to obtain embeddings with a 20 ms time resolution to capture dense, frame-level temporal dynamics; and inject this embedding into the backbone through an independent Temporal Feature-wise Linear Modulation (Temporal-FiLM) layer [7] to prevent information dilution.

*Corresponding Author

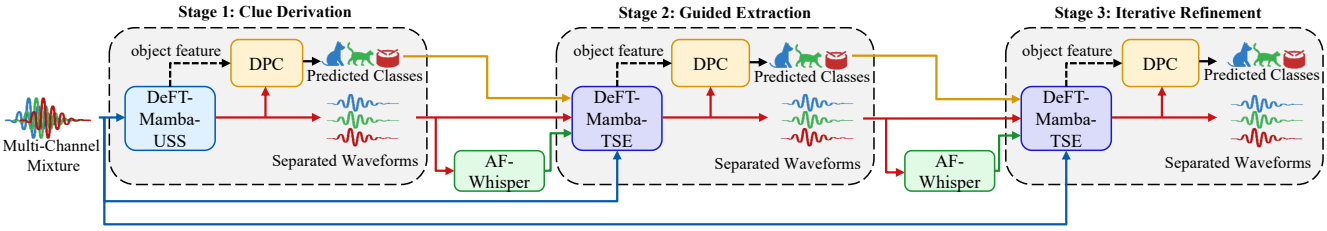

Figure 1: Overall architecture of the proposed multi-stage self-guided framework. Stage 1 derives initial clues using DeFT-Mamba-USS and M2D-DPC. Stages 2 and 3 iteratively refine target sound extraction using DeFT-Mamba-TSE, guided by the separated waveforms, one-hot predicted classes, and fine-grained AF-Whisper semantic embeddings.

- **Duration-based Augmentation** We introduce an augmentation strategy for difficult percussive classes by mixing short transient samples with longer sustained samples during training, thereby reducing the false-silence misclassification rate.
- **Class-specific Silence Threshold Optimization** To efficiently handle the silence prediction, we conduct a class-specific threshold tuning process to maximize the official metric, CAPI-SDRi.

In the official test set, our system achieves a CAPI-SDRi of 15.51 dB, a mixture accuracy of 71.09%, and a source accuracy of 78.62%, significantly outperforming the baseline by 7.02 dB, 10.37%p, and 8.23%p, respectively.

2. PROPOSED METHOD

2.1. Framework Overview

Figure 1 illustrates our proposed three-stage self-guided framework for joint separation and classification. In Stage 1 (Clue Derivation), DeFT-Mamba-USS decomposes the multi-channel mixture into distinct object features, which are directly processed by the DPC to predict class labels and detect silence. In Stage 2 (Guided Extraction) and Stage 3 (Iterative Refinement), DeFT-Mamba-TSE progressively refines the separated waveforms, followed by enhanced classification performance. Each refinement stage is guided by three complementary clues derived from the preceding stage: the separated waveform (enrollment clue), the predicted one-hot label (class clue), and a fine-grained semantic embedding extracted by AF-Whisper. The enrollment clues are concatenated with the multi-channel mixture along the channel dimension, and the class clues are injected through the FiLM [8] layers interleaved between the separation blocks.

2.2. AudioFlamingo-whisper Embedding

Although the features extracted by the pre-trained M2D model encode rich semantic information, they are derived from patchified representations with a temporal downsampling factor of 16, leading to relatively coarse temporal resolution and potentially limiting fine-grained temporal modeling. Therefore, we employ a fine-grained frame-level embedding from the AudioFlamingo (AF)-Whisper encoder [6], pretrained on a massive audio corpus.

As shown in Figure 2, this embedding is injected through additional Temporal-FiLM layers placed after each FiLM layer handling the one-hot vectors. This injection operates at a 20 ms time resolution to capture frame-level temporal dynamics. Let c_{oh} denote

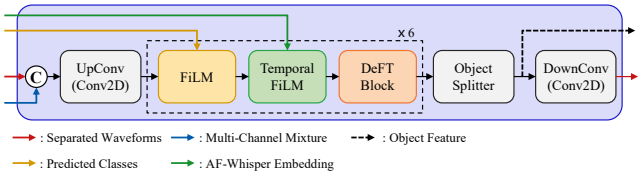

Figure 2: Architecture of the proposed DeFT-Mamba-TSE.

the embedding from the one-hot class vector predicted by the DPC, and E_{AF} represent the sequence of semantic audio embeddings extracted by AF-Whisper. To preserve definitive categorical information, c_{oh} is injected into the intermediate feature maps of DeFT-Mamba-TSE via a standard FiLM layer, providing global static conditioning. Immediately following this, the frame-level embedding sequence E_{AF} is injected through an additional Temporal-FiLM layer. This decoupled conditioning ensures that the TSE model benefits from dense temporal fluctuations for dynamic source tracking without compromising or overriding the strict categorical boundaries provided by the one-hot class clue.

2.3. Class-specific Silence Threshold Optimization

Since the acoustic characteristics vary between classes, the classification and silence detection performance also differ between classes. We observed that classes with highly percussive temporal characteristics (e.g., Percussion and Footsteps) or stationary acoustic patterns that resemble background noise (e.g., MechanicalFans) are more likely to be misclassified as silence. Furthermore, when the classifier predicts an incorrect class, the predicted silence score also tends to increase. To address this issue, we propose a class-wise silence threshold tuning strategy.

For a given test set, we apply a class-specific silence threshold that maximizes CAPI-SDRi. We first initialize the silence threshold of 0.5. Then sweep the threshold for each class and select the value that yields the highest CAPI-SDRi. The tuned thresholds are applied only during inference and are not used during training. The proposed method improves CAPI-SDRi in two ways. First, it improves the detection accuracy of zero-target events, thereby reducing the relative impact of CAPI-SDRi penalties. Second, when the classifier produces an incorrect class prediction, the tuned threshold increases the likelihood of assigning the output to silence instead. As a result, error cases that would otherwise incur both false-negative and false-positive penalties are converted into cases containing only a false-negative, leading to a higher CAPI-SDRi.

2.4. Duration-based Augmentation

The instability in detecting percussive sounds primarily stems from training instances containing only a single, brief transient event (e.g., an isolated drum hit), which lacks sufficient temporal context for the classification module to reliably distinguish it. Therefore, to mitigate the frequent misclassification of percussive sounds as silence, we propose a duration-based augmentation technique applied during the dynamic synthesis of the training data.

Specifically, we establish a duration threshold T_{th} for all samples belonging to percussive classes. During the batch generation process, instead of randomly sampling a single percussive audio file, we select two distinct files: a long sample x_{long} with an active sound duration equal to T_{th}, and a short sample x_{short} with a duration less than T_{th}. These two samples are stochastically mixed within the same spatial scene to form an augmented percussive source. By consistently exposing the network to percussive events with extended temporal presence, this augmentation drastically reduces the false-negative rate (silence misclassification) and enhances the extraction quality of short, impulsive acoustic events.

3. EXPERIMENTAL SETUPS

3.1. Datasets and Augmentation

To train our models, we dynamically generated 4-channel mixtures online for each sample using the SpatialScaper simulator [9]. For the duration-based augmentation, we set the threshold T_{th} to 4 seconds. While the foundational training data configuration follows the official DCASE 2026 Task 4 baseline [2], we introduced two modifications to improve data quality and inter-class discriminability:

- Speech: To ensure the model learns from high-quality human speech representations, we replaced the provided speech dataset with the high-fidelity VCTK corpus [10].
- Vacuum Cleaner: Preliminary experiments revealed that the baseline dataset did not provide sufficient acoustic variance to effectively distinguish the VacuumCleaner class from the Blender class. To resolve this ambiguity, we added a newly curated subset from AudioSet-2M [11] to the baseline vacuum cleaner data [2]. We strictly filtered this subset to include only high-quality samples annotated with a single "VacuumCleaner" label, minimizing label noise.

3.2. Loss Functions

Our multi-stage framework maintains a consistent loss function formulation across all training stages. At every stage, the network generates three distinct outputs for each estimated source: a separated waveform, a predicted class distribution (across the 18 target classes), and a scalar silence prediction. We optimize the network using a combined multi-task loss function defined as $\mathcal{L}_{total} = \mathcal{L}_{sep} + \mathcal{L}_{cls} + \mathcal{L}_{sil}$, where each component targets a specific output:

- Separation Loss ($\mathcal{L}_{sep}$): The separated waveforms are optimized using the Source-Aggregated Signal-to-Distortion Ratio (SA-SDR) loss [12] to ensure high-quality and phase-accurate source reconstruction.
- Classification Loss ($\mathcal{L}_{cls}$): The 18-class prediction is trained using the ArcFace loss [13]. This explicitly maximizes inter-class separability and intra-class compactness.
- Silence Prediction Loss ($\mathcal{L}_{sil}$): The scalar output indicating the presence or absence of an active source is optimized using Binary Cross-Entropy (BCE) loss.

3.3. Evaluation Metrics

To comprehensively evaluate both the separation quality and classification accuracy of our system, we employ three official metrics defined for DCASE 2026 Task 4: Class-Aware Permutation-Invariant Signal-to-Distortion Ratio improvement (CAPI-SDRi) [14], Mixture Accuracy ($\mathrm{Acc}_{\mathrm{mix}}$), and Source Accuracy ($\mathrm{Acc}_{\mathrm{src}}$).

CAPI-SDRi: As the primary ranking metric for this task, it extends the standard CA-SDRi to jointly evaluate sound event detection and separation, explicitly accommodating conditions where multiple sources of the same class co-occur. It utilizes a permutation-invariant objective to resolve assignment ambiguity for same-class sources. Under this metric, incorrect predictions (false positives and false negatives) strictly receive a 0 dB improvement penalty, while correct predictions contribute their permutation-optimized SDRi.

Mixture-level Accuracy ($\mathrm{Acc}_{\mathrm{mix}}$): This metric evaluates classification performance on a per-mixture basis. A prediction is considered correct only if the entire set of predicted labels for a mixture exactly matches the set of ground-truth labels.

Source-level Accuracy ($\mathrm{Acc}_{\mathrm{src}}$): This metric evaluates classification accuracy at the individual source level, measuring the proportion of correctly predicted labels among all separated foreground waveforms across the test set.

4. RESULTS AND DISCUSSIONS

4.1. Ablation Study

To validate the effectiveness of our proposed methods across the multi-stage framework, we conducted a comprehensive ablation study on the development test set. As shown in Table 1, our final proposed system (Stage 3 equipped with both AF-Whisper and threshold tuning) achieves a CAPI-SDRi of 15.51 dB and a mixture accuracy of 71.09%. This result significantly outperforms the official challenge baseline (CAPI-SDRi of 8.49 dB and $\mathrm{Acc}_{\mathrm{mix}}$ of 60.71%), proving the superiority of our framework.

The performance variations in Table 1 highlight several key findings. First, the **iterative refinement** process is highly effective even without the proposed modules, as demonstrated by the steadily improved base network's CAPI-SDRi from Stage 1 (11.05 dB) to Stage 3 (14.43 dB). Second, the **metric-driven threshold tuning** plays a crucial role in handling zero-target conditions. For instance, in Stage 1, applying this tuning strategy improves the CAPI-SDRi from 11.05 dB to 11.64 dB. Finally, the **fine-grained semantic conditioning via AF-Whisper** yields a substantial leap in separation quality. In Stage 3, the introduction of the AF-Whisper embedding increases the Source Accuracy from 75.63% to 76.09%, which demonstrates that the dense temporal resolution successfully guides the DeFT-Mamba-TSE to accurately track dynamic sources. In addition, when synergized with threshold tuning, the performance is further improved and reaches a peak CAPI-SDRi of 15.51 dB.

4.2. Class-specific Silence Threshold Optimization

When the silence threshold is optimized per class for CAPI-SDRi, we observe two distinct trends, illustrated by the two representative

Table 1: Ablation study on the development test set. The hyphen denotes that the module is not applicable or not used in that stage. CAPI-SDRi is reported in dB, while Mixture Accuracy ($\mathrm{Acc}_{\mathrm{mix}}$) and Source Accuracy ($\mathrm{Acc}_{\mathrm{src}}$) are reported in %.

Stage	AF-Whisper	Threshold Tuning	Development Test Set		
			CAPI-SDRi ↑	$\mathrm{Acc}_{\mathrm{mix}}$ ↑	$\mathrm{Acc}_{\mathrm{src}}$ ↑
baseline	-	-	8.49	60.71	70.39
1	-	✗	11.05	58.66	70.09
	-	✓	11.64	62.80	72.90
2	✗	✗	13.43	64.09	72.26
	✗	✓	13.72	66.01	72.47
	✓	✗	14.03	64.02	73.55
	✓	✓	14.26	65.48	74.44
3	✗	✗	14.43	65.41	75.63
	✗	✓	15.36	71.16	78.64
	✓	✗	14.65	66.07	76.09
	✓	✓	**15.51**	**71.09**	**78.62**

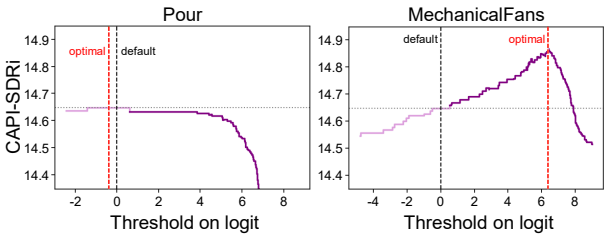

Figure 3: Impact of class-specific silence threshold tuning on CAPI-SDRi for the Pour (left) and Mechanical Fan (right) classes on the development test set.

classes in Figure 3. Note that the threshold on the x-axis represents the raw logit value. In the first trend (*Pour*, left), performance is largely insensitive to the threshold near the default (t=0, black dashed) and degrades only at large thresholds, so the optimal threshold (red dashed) yields little gain over the default. Conversely, for the second trend (*MechanicalFans*, right), performance is genuinely sensitive to the threshold and peaks at a substantially higher value, where the optimum clearly exceeds the default baseline.

For classes exhibiting this latter trend, a high threshold causes a larger fraction of outputs to be treated as silence, which is beneficial in two respects. First, when the model erroneously predicts an active source for a region that is in fact silent, a higher threshold correctly suppresses this output back to silence. Second, when the model assigns a region to an incorrect class, treating that output as silence removes a spurious detection (a false positive), avoiding the penalty it would otherwise incur in the CAPI-SDRi computation. Plotting curves analogous to Figure 3 for all classes, we find that *MechanicalFans*, *Clapping*, *Dishes*, *Footsteps*, and *Percussion* likewise require high thresholds.

4.3. Effect of Duration-based Augmentation

As observed during preliminary experiments, short and transient sound events are frequently missed by the classification module and incorrectly suppressed as silence. To explicitly evaluate the effectiveness of our duration-based augmentation in mitigating this issue, we analyzed the classification outcomes at the initial Clue Derivation stage (Stage 1) for three classes characterized by highly impulsive acoustics. Evaluating at Stage 1 is crucial because it serves as the foundational bottleneck of our multi-stage framework.

Table 2: Effect of duration-based augmentation on the classification outcomes for transient sound classes in the development test set. The values represent the number of samples from the initial Clue Derivation stage (Stage 1) predicted as the correct class, a wrong class, or misclassified as silence.

Class	Before Augmentation			After Augmentation		
	Correct	Wrong	Silence	Correct	Wrong	Silence
Percussion	47	19	74	76	36	28
Dishes	55	17	64	68	48	20
CupboardOpenClose	86	11	64	121	27	13

Table 2 compares the number of sample predictions before and after applying the augmentation strategy. For the *Percussion* class, the network initially struggled with brief transients, misclassifying 74 samples as silence. After implementing the duration-based augmentation, this false-negative count drastically dropped to 28, while correct predictions surged from 47 to 76. Similarly, for the *CupboardOpenClose* class, the augmentation resolved severe silence confusion, reducing 'Silence' predictions from 64 to 13 and nearly doubling the 'Correct' predictions (64 to 121). This detailed breakdown clearly demonstrates that our augmentation strategy reduces the false-silence rate for impulsive sounds.

5. CONCLUSIONS

In this report, we presented a multi-stage separation-and-classification framework for DCASE 2026 Task 4. To tackle the increased complexity of the new benchmark, specifically source-count uncertainty and zero-target events, we advanced our previous architecture by integrating a Dual-Path Classifier (DPC) that directly utilizes object features. Furthermore, we introduced fine-grained semantic conditioning via AF-Whisper to provide dense temporal guidance without diluting categorical boundaries, and a duration-based augmentation strategy to prevent the critical loss of transient sounds. Coupled with a metric-driven threshold tuning during inference, our fully equipped system achieved a CAPI-SDRi of 15.51 dB and a mixture accuracy of 71.09% on the development test set, demonstrating massive improvements over the challenge baseline. These results confirm that progressively refining internally generated multi-clues provides a highly robust and state-of-the-art solution for spatial semantic segmentation of sound scenes.

6. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337945), STEAM research grant (No. RS-2024-00464269) funded by the Ministry of Science and ICT of Korea government (MSIT), and the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE).

7. REFERENCES

[1] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, T. Nakatani, T. Kawamura, and N. Ono, "Description and discussion on dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes," 2025. [Online]. Available: https://arxiv.org/pdf/2506.10676v1

[2] B. T. Nguyen, M. Yasuda, N. Harada, R. Serizel, M. Mishra, M. Delcroix, C. Hernandez-Olivan, S. Araki, D. Takeuchi, T. Nakatani, and N. Ono, "Description and Discussion on DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes," 2026. [Online]. Available: https://arxiv.org/abs/2604.00776

[3] D. Lee, Y. Kwon, and J.-W. Choi, "DeepASA: An object-oriented multi-purpose network for auditory scene analysis," *Advances in Neural Information Processing Systems*, vol. 38, pp. 170 298–170 325, 2025.

[4] Y. Kwon, D. Lee, D. Kim, and J.-W. Choi, "Self-guided target sound extraction and classification through universal sound separation model and multiple clues," DCASE2025 Challenge, Tech. Rep., June 2025.

[5] Y. Kwon and J.-W. Choi, "Sound separation and classification with object and semantic guidance," *arXiv preprint arXiv:2509.15899*, 2025.

[6] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. Yang, R. Duraiswami, D. Manocha, R. Valle, *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *Advances in Neural Information Processing Systems*, vol. 38, pp. 41 819–41 886, 2026.

[7] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations." *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[9] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial scaper: a library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1221–1225.

[10] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[11] J. F. Gemmeke, D. P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[12] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Sa-sdr: A novel loss function for separation of meeting style data," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6022–6026.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[14] B. T. Nguyen, M. Yasuda, D. Takeuchi, D. Niizumi, and N. Harada, "Class-aware permutation-invariant signal-to-distortion ratio for semantic segmentation of sound scene with same-class sources," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 15 862–15 866.