forum acusticum 2026
8-12 September · Graz, Austria

SOUND SEPARATION AND CLASSIFICATION WITH OBJECT AND SEMANTIC GUIDANCE

Younghoo Kwon[1], Jung-Woo Choi[1*]

[1] *School of Electrical Engineering, KAIST, Korea*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

The spatial semantic segmentation task focuses on separating and classifying sound objects from multichannel signals. To achieve two different goals, conventional methods fine-tune a large classification model cascaded with the separation model and inject classified labels as separation clues for the next iteration step. However, such integration is not ideal, as fine-tuning over a smaller dataset loses the diversity of large classification models; features from the source separation model are different from the inputs of the pretrained classifier, and injected one-hot class labels lack semantic depth, often leading to error propagation. To resolve these issues, we propose a Dual-Path Classifier (DPC) architecture that combines object features from a source separation model with semantic representations acquired from a pretrained classification model without fine-tuning. We also introduce Semantic Embedding Conditioning (SEC), which enriches the semantic depth of injected clues. Our system achieves a state-of-the-art 11.19 dB CA-SDRi and enhanced semantic fidelity on the DCASE 2025 task4 evaluation set, surpassing the top-ranked performance of 11.00 dB. These results highlight the effectiveness of integrating separator-derived features and rich semantic clues.

Keywords: *target sound extraction, sound classification, dual-path classifier, semantic clue encoder*

1. Introduction

Comprehending auditory scenes is crucial in machine listening to interpret and engage with complex soundscapes. Separation and classification of individual sound sources are two key tasks to achieve this goal. While the sound separation performance has improved with advances in the models [1], [2], conventional separation models are limited by the maximum number of sources they can process. To overcome this limitation, researchers have developed methods that can extract a specific sound from the complex sound scene based on a user-provided clue. This line of research, named target signal extraction (TSE), has been developed to process various types of clues, such as a given class label [3], a similar audio sample [4], [5], or a timestamp [6]. Meanwhile, classifiers [7], [8] have evolved to utilize features from the self-supervised models [9], [10] pretrained on a large-scale dataset [11]. For joint separation and classification, however, various approaches are possible, e.g., first performing multi-label classification on the mixture and using the resulting class predictions as clues for target sound extraction [12], [13]. Another strategy, embodied by DeFT-Mamba [14], is first to separate each source from the mixture and then classify them.

The DCASE 2025 task4 challenge, known as Spatial Semantic Segmentation of Sound Scenes (S5) [15], [16], has been designed to reflect the growing focus of research. The challenge focuses on separating and classifying foreground sound sources from multichannel mixtures, which also include interfering sounds and background noise. Each mixture can contain up to three foreground sources, and each foreground signal is categorized into one of the 18 designated classes.

Our previous work, which achieved the 1st rank in the challenge [17], employs a three-stage framework. As depicted in Fig. 1 (a), the waveforms separated in the previous stage and the classes estimated from those separated waveforms are used as clues for foreground sound extraction in the next stage. Across the entire framework, three models are employed. First, DeFT-Mamba [14] for Universal Sound Separation (DM-USS) performs separation without any clue. DM-USS separates each sound source into rich object features, which are decoded into the separated waveform through a weight-shared audio decoder, as shown in Fig. 1 (b). Second, DeFT-Mamba for Target Sound Extraction (DM-TSE) extracts foreground sounds from the mixture using enrollment clues and class clues. The enrollment clues are concatenated with the multichannel

***Corresponding author**: jwoo@kaist.ac.kr.*

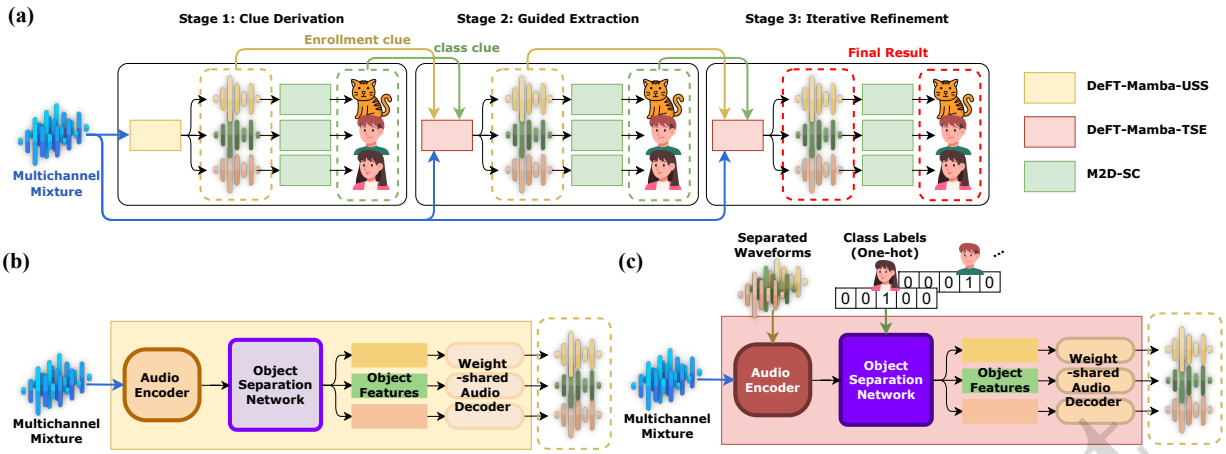

Figure 1: Architecture of the previous multi-stage framework used in our DCASE challenge system. (a) Overall framework. (b) Detailed architecture of DM-USS. (c) Detailed architecture of DM-TSE.

mixture along the channel dimension, providing direct separation clues. The class clues are injected in the form of one-hot vectors through Residual FiLM [18], [19]. DM-TSE follows the same architecture, except for these clue injection modules as seen in Fig. 1 (c). Finally, M2D [10] for Single-label Classification (M2D-SC) classifies a single class from each separated waveform and is used as the classifier throughout all stages.

Based on these three models, the overall framework proceeds in three stages. In Stage 1 (**clue derivation**), DM-USS performs separation on the input mixture, producing separated waveforms which are further classified by M2D-SC to estimate a single class for each separated sound. In Stage 2 (**guided extraction**), DM-TSE extracts foreground sounds from the mixture using the separated waveforms from the previous stage as enrollment clues and the estimated classes as class clues. In Stage 3 (**Iterative refinement**), the extraction process is repeated by reusing the outputs of the previous stage as updated clues, thereby progressively refining the extracted foreground sounds.

The current design exhibits several limitations due to the coarse combination with the existing classification model and the naive clue injection module in DM-TSE. (1) The M2D-SC is based on the M2D pretrained over the large corpus [11], but fine-tuning the pretrained M2D on the smaller S5 dataset [15] compresses the pretrained diversity [20], [21]. (2) Separated object features are converted into Mel-spectrograms to make them compatible with M2D. However, the magnitude-only projections with spectral merging lead to a significant loss of crucial phase and fine-grained frequency information. (3) The class embeddings based on one-hot vectors fail to encode rich semantic information. (4) The class-based guidance for the TSE model depends solely on the predicted label so that an early misclassification can steer DM-TSE in the wrong direction throughout the extraction stages. Therefore, we need a novel classification architecture that can leverage the knowledge obtained from the pretraining and inject semantic clues with sufficient depth.

We propose a Dual-Path Classifier (DPC) to address the lack of diversity in the M2D-SC and the informative insufficiency of the Mel-spectrogram. We adopt the lightweight CRNN instead of fine-tuning M2D to preserve the diversity of the features from M2D. The object features of the separators are fed into the CRNN instead of the Mel-spectrogram. Unlike Mel-spectrogram, these features retain crucial phase and fine-grained frequency information, providing a richer input for the classifier. Additionally, we propose a Semantic Embedding Conditioning (SEC) to overcome the static nature of the one-hot condition embedding and the critical flow of misclassification. The SEC solves this issue by supplementing the conventional one-hot condition embedding with richer embeddings from models trained with self-supervised methods [10] or contrastive learning [22]. The injection of these additional semantic clues through the SEC makes the DM-TSE less critically dependent on the label estimation of the previous stage, thereby enhancing the overall robustness of the system. Eventually, our system achieves a final CA-SDRi of 11.19 dB on the evaluation dataset and 16.60 dB on the test dataset [16], corresponding to gains of +0.19 dB and +1.66 dB, respectively, over our previous submission.

2. Methodology and architecture

The proposed architecture applies DPC and enhances the DM-TSE model with the SEC using the semantic condition embeddings. Two individual instances of the DPC are trained: one for DM-USS (DPC-USS) and another for DM-TSE (DPC-TSE).

2.1. Dual-Path Classifier (DPC)

The proposed DPC is detailed in Fig. 2. The core of the DPC is a dual-path CRNN that operates on the object features of the separators. The CRNN consists of symmetric temporal and frequency paths (T/F-paths), each containing a sequence of a CNN, a Feature Fusion (FF) block, a GRU [23], and an ASP [24] layer. The feature from each path is fused

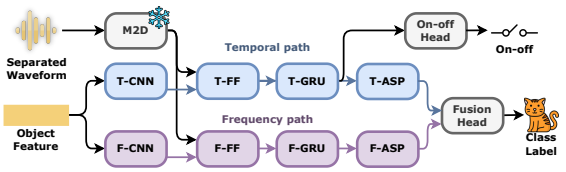

Figure 2: The proposed pipeline of the Dual-Path Classifier for the object features.

to produce logits in the fusion head. Additionally, the on-off head determines whether the separated source is active or silent.

T/F-CNN: A stack of 2D convolutions and pooling layers compresses the object feature $\mathbf{X} \in \mathbb{R}^{D\times F\times T}$ along the time and frequency axes, where D, F, and T denote the channel, frequency, and time dimension, respectively. The output of CNN in the temporal path is $\mathbf{X}_t \in \mathbb{R}^{D\times F_t\times T_t}$, while the frequency path employs a final point-wise convolution that doubles the channel dimension, resulting in $\mathbf{X}_f \in \mathbb{R}^{2D\times F_f\times T_f}$. The F and T with the subscript represent the compressed frequency and time resolutions, where the subscripts f and t denote the frequency and temporal paths, respectively. The features are then reshaped for sequential processing. The temporal feature $\mathbf{X}_t$ is flattened to preserve the time axis, yielding $\mathbf{X}_t \in \mathbb{R}^{T_t\times DF_t}$, while the frequency feature $\mathbf{X}_f$ is averaged over the time dimension, treating the frequency axis as a sequence dimension, resulting $\mathbf{X}_f \in \mathbb{R}^{F_f\times 2D}$.

Feature Fusion (FF) block and GRU: The FF block aggregates the features of M2D and CNN in each path. A feature $\mathbf{X}_{\text{M2D}} \in \mathbb{R}^{G\times F_{\text{M2D}}\times T_{\text{M2D}}}$ is extracted from the M2D to leverage knowledge from a pretrained model, where G, F_{M2D}, and T_{M2D} denote the channel, frequency, and time dimension of the feature from M2D. To prepare for fusion, the $\mathbf{X}_{\text{M2D}}$ is reshaped differently for each path to match the dimensions of the corresponding CNN feature. For the temporal path, $\mathbf{X}_{\text{M2D}}$ is first average-pooled along the frequency axis and then upsampled via bilinear interpolation along the time axis to match the sequence length of $\mathbf{X}_t$, resulting $\mathbf{X}_{\text{M2D},T} \in \mathbb{R}^{G\times T_t}$. For the frequency path, a symmetric operation is performed to produce $\mathbf{X}_{\text{M2D},F} \in \mathbb{R}^{G\times F_f}$. Each aligned M2D feature is then concatenated with its corresponding sequence feature from the CNN along the embedding dimension. These concatenated features are then passed through the FF block, which is the linear layer. This unified feature sequence is then fed into a bidirectional GRU, which models long-range contextual dependencies.

ASP and Fusion head: The ASP layer converts the sequential feature of each GRU into a fixed-size embedding vector. Specifically, each branch contains its own ASP layer that processes the respective GRU output, creating two distinct embedding vectors. The embeddings from the temporal and frequency paths are concatenated and passed through a fusion head which is a linear layer to produce the logits.

On-off prediction: The temporal path also handles on-off (silence) prediction using the on-off head on the sequential output of GRU in the temporal path. This head produces frame-wise activity probabilities with a linear layer and sigmoid function, and the source is classified as 'on' if the maximum probability across frames exceeds a threshold of 0.5.

2.2. Semantic Embedding Conditioning (SEC)

Semantic Embedding Conditioning (SEC) is a conditioning strategy that enriches the original one-hot label class conditioning [17], [25] with semantically informative embeddings from pretrained models. The label-based conditioning has two limitations which are the lack of semantic richness and error propagation. To address these limitations, SEC incorporates semantic embeddings obtained from M2D [10] and MGA-CLAP [22]. The semantic embedding is projected to the same dimensionality as the one-hot embedding, and the two embeddings are added element-wise. The combined embeddings are finally injected into DM-TSE thorugh the same Residual FiLM [18]. In this way, the resulting condition embeddings alleviate the error propagation caused by hard label conditioning by providing semantically informative cues.

3. EXPERIMENTS

3.1. Datasets & Evaluation metrics

The training data was generated dynamically for each sample. We synthesized 4-channel mixtures on-the-fly using the SpatialScaper synthesis approach [26]. Additional data was incorporated for training. The speech class was supplemented with the VCTK corpus [27], and the percussion class was enriched with extra samples from open-source databases [17]. For evaluation, we reported results on the official DCASE 2025 task4 test and evaluation datasets. The test dataset was provided for the intermediate assessment during the challenge, while the evaluation dataset was used for the final official ranking of the challenge submissions. The 1,620-sample subset of the evaluation data was used for the official challenge ranking.

We evaluated the performance of our systems using four metrics. The Class-Aware Signal-to-Distortion Ratio improvement (CA-SDRi) [15], [16] is the primary metric. CA-SDRi is based on the standard SDRi, a measure of separation quality, but is designed to penalize incorrect classifications. If a source is misclassified, its SDRi is disregarded. The mixture-level accuracy (Acc_{mix}) [25] is used to evaluate classification performace. Acc_{mix} is calculated on a per-mixture basis, where a single mixture is considered correct only if the classifications for all of its constituent foregrounds are correct.

3.2. Experimental settings

Our experiments were based on the pretrained weights of the previous system [17] on the challenge. To isolate the performance gains of our contributions, we froze the parameters of the core separation models (DM-USS and DM-TSE backbone). The only trainable modules in our experiments were the proposed DPC and SEC. After training the DPC, we froze its

Table 1: Ablation study on the DPC and SEC across each stage. The hyphen in the 'Semantic embedding' column denotes conditioning solely on the one-hot condition embedding without any semantic information. CA-SDRi in [dB] and Acc_{mix} in [%].

Stage	Classifier	Semantic embedding	Eval dataset		Test dataset	
			CA-SDRi ↑	Acc_{mix} ↑	CA-SDRi ↑	Acc_{mix} ↑
1	M2D-SC	-	8.47	52.22	12.70	59.07
	DPC	-	8.18	52.04	12.81	69.20
2	M2D-SC	-	10.08	54.01	14.70	61.26
	DPC	-	10.86	57.78	16.21	72.13
		MGA-CLAP-text	10.97	58.27	16.51	73.73
		MGA-CLAP-audio	10.97	58.89	16.50	73.53
		M2D	10.94	57.90	16.44	73.40
3	M2D-SC	-	11.00	55.80	14.94	61.80
	DPC	-	11.06	58.64	16.34	72.93
		MGA-CLAP-text	11.18	59.38	**16.60**	74.20
		MGA-CLAP-audio	11.11	59.01	16.56	**74.33**
		M2D	**11.19**	**59.63**	**16.60**	74.20

parameters and then trained the SEC separately using three distinct semantic embeddings: those from M2D and from the text and audio encoders of MGA-CLAP. AdamW optimizer was used for training both DPC and SEC. The learning rates for DPC and SEC were 4×10^{-4} and 4×10^{-6}, respectively. Cross-entropy and KL divergence loss functions were used to train the classification ability of the DPC for active foreground and silence, respectively. Binary cross-entropy was utilized to detect silence. The masked SNR loss function was used for training the SEC. The detailed settings for the loss functions follow [25].

4. Results

4.1. Ablation Study on DPC and SEC

Table 1 compares the classifiers (M2D-SC and DPC) and semantic clues across the three stages. We report performance on both the evaluation and test datasets, considering the evaluation dataset a more robust indicator of generalization.

Regarding the classifier performance, the proposed DPC consistently outperforms the conventional M2D-SC in Acc_{mix} in Stage 2 and 3 across all datasets. The improvements are particularly substantial on the test dataset, where DPC achieves accuracy gains of over 10%p in every stage. Specifically, the accuracy increases from 59.07% to 69.20% (+10.13%p) in Stage 1, from 61.26% to 72.13% (+10.87%p) in Stage 2, and from 61.80% to 72.93% (+11.13%p) in Stage 3. Notably, the margin of improvement for DPC over M2D-SC widens as the separation quality improves in the later stages. This result suggests that DPC, by directly utilizing the object feature of the separators, is more sensitive and responsive to the separators' performance. Furthermore, the DPC also demonstrates strong generalization. On the evaluation dataset, DPC achieves a 2.84%p accuracy improvement in Stage 3. We attribute this generalization to the absence of fine-tuning on the M2D, which allows our systems to retain diverse features from its pretraining.

The results in Table 1 clearly demonstrate that incorporating the semantic embeddings consistently improves performance over conditioning only on the one-hot vectors, particularly for the CA-SDRi. Especially, the improvement in the evaluation dataset from 11.06 dB to 11.19 dB (+0.13 dB) represents that the benefits of the SEC generalize well to unseen data. Ultimately, these results indicate that the SEC not only enhances the objective separation and classification quality (CA-SDRi) but also holistically improve the system's output, leading to more accurate classifications and audio that is perceptually and semantically more faithful to the target source.

Table 2: DCASE 2025 Task 4 leaderboard with CA-SDRi in [dB] and Acc_{mix} in [%].

Model	Evaluation Set		Complexity	
	CA-SDRi ↑	Acc_{mix} ↑	Ensenble	Param
Proposed	**11.19**	<u>59.63</u>	1	205M
Rank 1 [25]	<u>11.00</u>	55.80	1	204M
Rank 2 [28]	9.77	**61.60**	118	10.8B
w/o ensemble	8.91	51.98	1	229M
Rank 3 [29]	9.73	51.54	1	87M
Rank 4 [30]	7.84	47.72	1	105M
Rank 5 [31]	7.55	49.51	10	25M
Rank 6 [32]	6.69	47.22	4	465M
w/o ensemble	6.64	46.67	1	116M
Rank 7 [33]	6.60	51.48	1	115M
Baseline [16]	6.60	51.48	1	115M
Rank 8 [34]	3.84	22.41	1	115M

4.2. Comparison with Challenge Systems

Our proposed system, which employs the SEC with M2D embeddings, achieves the best CA-SDRi of 11.19 dB among all systems, while also attaining a strong classification accuracy of 59.63%. Although the second-ranked team achieves the highest Acc_{mix} of 61.60%, this result is obtained using an ensemble of 118 models, resulting in a substantially increased model complexity of 10.82B parameters. In contrast, the corresponding non-ensemble submission of the same team achieves only 51.98% in Acc_{mix}. These results indicate that our DPC-based framework remains highly effective for classification improvement, achieving competitive classification performance without relying on large-scale ensembling, while simultaneously providing the best separation performance.

4.3. Classification error correction by SEC

Table 3 evaluates the error correction capacities of SEC in mitigating error propagation. The table categorizes samples based on their transition of classification correctness from Stage 1 to Stage 2: Wrong-to-Correct (W2C) for corrected errors, Correct-to-Correct (C2C) for stable predictions, and Correct-to-Wrong (C2W) for newly introduced errors. For each category, we report the ratio of samples relative to the entire dataset and the average improvement in separation quality (ΔCA-SDRi).

On the test dataset, using the SEC leads to comprehensive improvements across all categories for both the sample ratios and the ΔCA-SDRi. These results

Table 3: Comparison of error correction performance by embedding type.

Dataset	Embedding type	Ratio of samples (%)			Δ CA-SDRi (dB)		
		W2C ↑	C2C ↑	C2W ↓	W2C ↑	C2C ↑	C2W ↑
Test	one-hot vector	10.5	61.7	7.5	11.75	3.46	-5.71
	+ MGA-CLAP-text	11.3	**62.4**	**6.8**	11.78	**3.67**	-5.63
	+ MGA-CLAP-audio	11.3	62.3	6.9	**11.86**	3.64	**-5.27**
	+ M2D	**11.4**	62.0	7.2	11.82	3.64	-5.71
Eval	one-hot vector	13.5	**44.3**	**7.8**	9.89	2.43	-4.60
	+ MGA-CLAP-text	14.5	43.8	8.3	**10.06**	2.52	-4.69
	+ MGA-CLAP-audio	**14.8**	44.1	7.9	10.04	2.51	-4.83
	+ M2D	14.4	43.5	8.6	10.01	**2.53**	**-4.48**

suggest that the SEC not only mitigates error propagation by improving classification transitions but also enhances the separation performance. Upon closer examination, different semantic embeddings exhibit distinct strengths. The M2D embedding proves to be the most effective at error correction, achieving the highest W2C ratio of 11.4%. Meanwhile, the MGA-CLAP-text embedding excels at preserving correct predictions from the previous stage, evidenced by the highest C2C ratio of 62.4% and the lowest C2W ratio of 6.8%. Furthermore, the MGA-CLAP-audio shows the largest ΔCA-SDRi of 11.86 dB for W2C and the smallest degradation of -5.27 dB for C2W, thereby boosting the overall performance.

On the evaluation dataset, using only the one-hot vector shows slightly superior prediction preservation in terms of sample ratio. For example, the MGA-CLAP-audio embedding reduces the C2C ratio from 44.3% to 44.1%. Despite this degradation, the benefit of the SEC becomes clear when observing the ΔCA-SDRi. For W2C samples, all semantic embeddings outperform the one-hot vector by more than 0.1 dB. For the other categories, the M2D embedding demonstrates an improvement of 0.10 dB from 2.43 dB to 2.53 dB on the C2C and 0.12 dB from -4.60 dB to -4.48 dB on the C2W. Collectively, the results in Table 3 confirm that the use of the SEC with additional semantic embeddings improves the final CA-SDRi performance.

5. Conclusion

In this paper, we addressed two key limitations of the first-place system from DCASE 2025 Task 4: an inefficiently integrated classifier architecture and a weak conditioning mechanism susceptible to error propagation. To resolve these issues, we introduced the Dual-Path Classifier (DPC), which operates on object features from the separator and semantic features from a pretrained model. We also proposed the Semantic Embedding Conditioning (SEC) to enrich the conditioning signal with content-based embeddings from large-scale models. Our experiments demonstrate that the DPC significantly outperforms the baseline classifier and that the SEC provides substantial additional gains, achieving a state-of-the-art CA-SDRi of 11.19 dB on the evaluation set. Furthermore, our analysis revealed that the SEC enhances system robustness through a dual mechanism: the SEC not only corrects a greater number of initial classification errors (W2C), but also improves the separation quality with the rich semantic information.

6. Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337945) and STEAM research grant (No. RS-2024-00464269) funded by the Ministry of Science and ICT of Korea government (MSIT), the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE).This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

[1] D. Lee and J.-W. Choi, "DeFTAN-II: Efficient multichannel speech enhancement with subgroup processing," *TASLP*, vol. 32, pp. 4850–4866, 2024.

[2] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Real-time target sound extraction," in *Proc. ICASSP*, 2023, pp. 1–5.

[4] 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," *TASLP*, vol. 31, pp. 121–136, 2023.

[5] C. Hernandez-Olivan et al., "SoundBeam meets M2D: Target sound extraction with audio foundation model," in *Proc. ICASSP*, 2025, pp. 1–5.

[6] D. Kim, M.-S. Baek, Y. Kim, and J.-H. Chang, "Improving target sound extraction with timestamp knowledge distillation," in *Proc. ICASSP*, 2024, pp. 1396–1400.

[7] F. Schmid, P. Primus, T. Morocutti, J. Greif, and G. Widmer, "Improving audio spectrogram transformers for sound event detection through multi-stage training," DCASE2024 Challenge, Tech. Rep., May 2024.

[8] H. Nam, D. Min, S. Choi, I. Choi, and Y.-H. Park, "Self training and ensembling frequency dependent networks with coarse prediction pooling and sound event bounding boxes," DCASE2024 Challenge, Tech. Rep., May 2024.

[9] X. LI and X. Li, "ATST: Audio Representation Learning with Teacher-Student Transformer," in *Proc. Interspeech*, 2022, 4172–4176. DOI: {10.21437/Interspeech.2022-10126}.

[10] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *TASLP*, vol. 32, pp. 2391–2406, 2024.

[11] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, IEEE, 2017, pp. 776–780.

[12] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *Proc. ICASSP*, 2020, pp. 96–100.

[13] Y. Liu et al., "Audio prompt tuning for universal sound separation," in *Proc. ICASSP*, 2024, pp. 1446–1450.

[14] D. Lee and J.-W. Choi, "DeFT-Mamba: Universal multichannel sound separation and polyphonic audio classification," in *Proc. ICASSP*, 2025, pp. 1–5.

[15] M. Yasuda et al., *Description and discussion on dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes*, 2025. arXiv: 2506.10676v1 [eess.AS].

[16] 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*, 2025. arXiv: 2503.22088 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2503.22088.

[17] 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., Jun. 2025.

[18] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, vol. 32, 2018.

[19] Q. Kong et al., "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.

[20] S. Yamaguchi, S. Kanai, K. Adachi, and D. Chijiwa, "Adaptive random feature regularization on fine-tuning deep neural networks," in *Proc.CVPR*, Jun. 2024, pp. 23 481–23 490.

[21] R. Perera and S. Halgamuge, "Discriminative sample-guided and parameter-efficient feature space adaptation for cross-domain few-shot learning," in *Proc.CVPR*, Jun. 2024, pp. 23 794–23 804.

[22] Y. Li, Z. Guo, X. Wang, and H. Liu, "Advancing multi-grained alignment for contrastive language-audio pre-training," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7356–7365.

[23] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 2017, pp. 1597–1600.

[24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[25] 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," in *DCASE*, 2025.

[26] 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 *Proc. ICASSP*, 2024, pp. 1221–1225.

[27] C. Veaux, J. Yamagishi, and K. MacDonald, *CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit*, 2017.

[28] T. Morocutti, F. Schmid, J. Greif, P. Primus, and G. Widmer, "Transformer-aided audio source separation with temporal guidance and iterative refinement," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[29] F. Wu and Z.-Q. Wang, "TS-TFGRIDNET: Extending tfgridnet for label-queried target sound extraction via embedding concatentaiton," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[30] X. Zhou, H. Wang, C. Li, B. Han, X. Zheng, and Y. Qian, "Sjtu-audiocc system for dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[31] Y. Nozaki, S. Sakurai, Y. Bando, K. Saijo, K. Imoto, and M. Onishi, "A hybrid s5 system based on neural blind source separation," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[32] J. Park, J. Lee, D.-H. Lim, H. K. Kim, H. Geum, and J. E. Lim, "Performance improvement of spatial semantic segmentation with enriched audio features and agent-based error correction for dcase 2025 challenge task 4," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[33] V. Stergioulis and G. Potamianos, "Redux: An iterative strategy for semantic source separation," DCASE2025 Challenge, Tech. Rep., Jun. 2025.

[34] Z. Wang, S. Wang, Z. Zhang, and J. Yin, "Spatial semantic segmentation of sound scenes based on adapter fine-tuning," DCASE2025 Challenge, Tech. Rep., Jun. 2025.