Local-Global Transformer with Iterative Refinement for Multi-Channel Sound Source Separation and Extraction

Technical Report

Ruohan Wang[1], Minjun Chen[1], Yangyang Liu[1], Longhai Wu[1], Jie Chen[1]
[1] Samsung Research China-Nanjing, Nanjing, China
{ruohan1.wang, minjun.chen, yang17.liu, longhai.wu, ada.chen}@samsung.com

ABSTRACT

This technical report describes our proposed systems for DCASE2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). The task aims to enhance technologies for sound event detection and separation from multi-channel input signals that mix multiple sound events with spatial information, which is a fundamental basis of immersive communication. Our approach employs a deep frequency-time transformer architecture with local modeling by convolution that processes multi-channel audio recordings from a 4-microphone array. The system consists of three main components: (1) a universal sound separation module for separating waveform and predicting the sound classes at the same time, (2) an audio tagging model for semantic label prediction, and (3) an iterative target sound extraction module that leverages enrollment clues and semantic labels to extract specific sound sources. We incorporate spatial features including inter-channel phase difference (IPD) and inter-channel level difference (ILD) to enhance separation performance.

Index Terms— Sound source separation, target sound extraction, transformer networks, audio tagging

1. INTRODUCTION

The DCASE2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5) aims to enhance technologies for sound event detection and separation from multi-channel input signals that mix multiple sound events with spatial information[1]. This task requires systems to identify, localize, and separate multiple sound sources in complex acoustic environments. The S5 task addresses technologies that form a fundamental basis of immersive communication, where systems must handle overlapping sound sources from the same class and correctly process mixtures that may contain zero target events.

Our approach addresses these challenges through a three-stage pipeline that progressively refines sound source representations. The first stage implements the universal sound separation (USS) model to separate foreground sound events, interference, and noise signals. The second stage employs an audio tagging (AT) model to predict semantic labels for sound events. The third stage performs iterative target sound extraction (TSE) using enrollment clues and predicted semantic labels to extract specific sound sources of interest. Spatial features are computed from multi-channel complex spectra: IPD (Inter-channel Phase Difference) represents phase difference between each microphone and the reference microphone, normalized to [-1, 1]. ILD (Inter-channel Level Difference) represents log-magnitude difference between each microphone and the reference, clipped to [-10, 10] dB and normalized. The combined spatial feature vector has dimension 2(M-1) = 6 for M=4 microphones.

The core of our architecture is built upon transformer networks with axial attention mechanisms that efficiently process frequency-time representations. The model uses feedforward networks (FFNs) with convolution layers, instead of linear layers, to capture local information, allowing self-attention to focus on capturing global patterns. We place two such FFNs before and after self-attention to enhance local-modeling capability. Unlike recurrent architectures, transformers enable parallel processing and capture long-range dependencies in both frequency and time dimensions. We incorporate spatial features derived from multi-channel recordings to leverage directional information for improved separation.

2. SYSTEM ARCHITECTURE

The proposed system processes 4-channel first-order Ambisonics audio recordings through a three-stage cascade pipeline, as illustrated in Figure 1. Given a multi-channel input x ∈ R^(M× T), where M = 4 is the number of microphone channels and T is the number of time samples, the system progressively refines sound source representations through the following stages. In Stage 1, Universal Sound Separation (USS) employs the TF-Locoformer architecture[2] to separate the mixture into 3 foreground sources, 2 interference sources, and 1 noise component, using class-aware permutation invariant training (CA-PIT)[3] to handle same-class source ambiguity. In Stage 2, the Source Classifier (SC) predicts semantic class labels for the separated foreground sources by fusing a fine-tuned M2D [4]audio backbone with three frozen pretrained SED encoders (BEATs, ATST-F, fPaSST)[5] through a learnable fusion module, and additionally performs silence detection via energy-based thresholding. In Stage 3, Target Sound Extraction (TSE) iteratively extracts specific target sources using dual conditioning signals—enrollment audio concatenated along the input channel dimension and class clues applied through the FiLM [6] mechanism—with each iteration's output serving as enrollment for the next refinement step. All modules operate on time-frequency representations computed via STFT with 1024-point FFT size and 160-sample hop length (5 ms at 32 kHz sampling rate), and incorporate spatial features including inter-channel phase difference (IPD) and inter-channel level difference (ILD) derived from the multi-channel complex spectra.

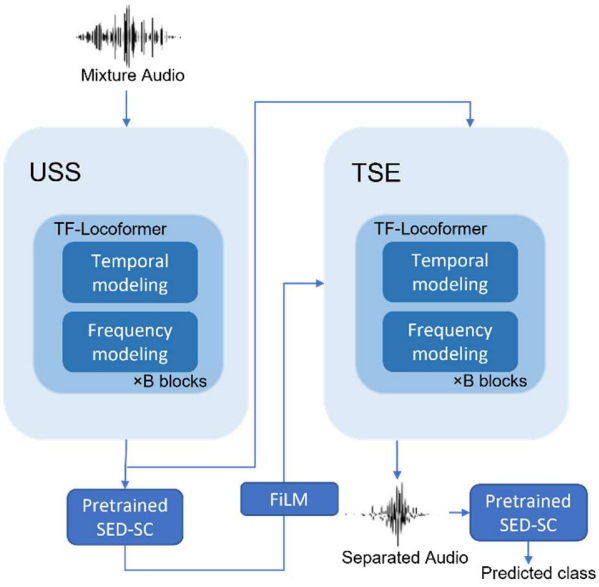

Figure 1: Overall framework of the proposed method

2.1. Separation architecture

Figure 2 shows the overview of the TF-Locoformer structure. The core architecture for separation employs TF-Locoformer (TF-domain Transformer with Local modeling by Convolution), a simple extension of the Transformer-based model that alternately performs global and local modeling. Starting from the standard Transformer block, we first replace the two linear layers in the feed-forward network (FFN) with 1D convolutional layers and 1D deconvolutional layers, respectively. Furthermore, we utilize SwiGLU [7][8] activations in the FFN and, drawing on the successful experience of the macaron-style architecture, place such FFNs before and after self-attention.

2.2. UNIVERSAL SOUND SEPARATION

The USS model employs the TF-Locoformer architecture described above, designing three output heads to meet the dual requirements of separation and classification. The mask head uses 1x1 convolutions to map 128 dimensional features into 3 channels of time-frequency masks, supporting both sigmoid activation (for independent masks) and SoftMax activation (for competitive masks). The generated masks are restored to the original STFT resolution via bilinear interpolation to achieve precise frequency-domain reconstruction. The Class Head utilizes adaptive average pooling and fully connected layers to predict 18+1-dimensional class logits (including the silence class) for each output slot, realizing the semantic classification of sound sources. The Silence Head, similarly based on global pooling and linear projection, predicts a silence probability logit for each output slot to detect inactive source segments.

2.3. TARGET SOUND EXTRACTION

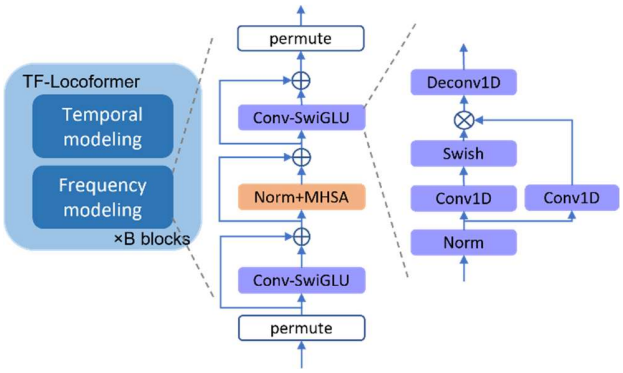

Figure 2: Overview of the TF-Locoformer

The TSE module serves as the second stage of the two-phase separation pipeline, responsible for extracting specific target sound sources from mixed audio. Unlike the universal sound separation in the USS stage, the TSE stage utilizes two types of conditioning signals—enrollment audio and class clues—to guide the model in extracting the designated target sound sources. The core processing module of TSE shares the same TFLocoformer architecture with USS. Enrollment audio is injected into the model by concatenating along the input channel dimension, while class clues are applied through the FiLM mechanism to exert conditional influence in the deep feature space. The enrollment audio provides low-level acoustic matching information, and the class clues offer high-level semantic guidance—enabling TSE to accurately locate and extract target sound sources within complex acoustic environments.

The TSE model performs iterative extraction, where each iteration refines the target estimate using information from previous iterations. We use 1 or 2 iterations in our proposed system. At each iteration, the current target estimate serves as enrollment for the next iteration, the USS foreground outputs provide initial source estimates, and the TSE model combines enrollment and class clues to extract the target.

2.4. AUDIO TAGGING MODEL

The Source Classifier (SC) model is a multi-branch audio classification architecture designed for single-label source identification of separated waveforms. It fuses a fine-tuned Masked Modeling Duo (M2D) audio backbone with three frozen PretrainedSED encoders (BEATs, ATST-F, fPaSST) [4] through a learnable fusion module, and additionally predicts frame-level activity to guide activity-aware temporal pooling. The model outputs class predictions with silence detection via energy-based thresholding, and serves as a frozen teacher in the TSE training pipeline to provide soft class labels.

M2D Backbone Branch: The primary branch employs a PortableM2D (ViT-Base) encoder to extract frame-level features, predicts per-frame activity logits via an activity head, and performs activity-weighted temporal pooling by concatenating four statistics: global mean, global max, weighted mean, and weighted standard deviation. The pooled features are projected to a 512-dimensional embedding via an MLP, with only the last 4 Transformer blocks and projection heads unfrozen for fine-tuning.

PretrainedSED Branches: Three frozen pretrained audio encoders from the PretrainedSED repository serve as auxiliary

semantic branches. BEATs[9]: Microsoft's audio Transformer, loaded with the "strong_1" checkpoint (trained on strongly-labeled audio events). Outputs 768-dimensional embeddings. ATST-F[10]: Audio-level self-supervised Transformer, also loaded with "strong_1" checkpoint. Outputs 768-dimensional embeddings. fPaSST[11]: Frame-level Patchout audio Transformer, loaded with "strong_1" checkpoint. Outputs 768-dimensional embeddings.

At inference, silence is detected via the energy score E = −logsumexp(plain_logits), with per-class calibrated thresholds; labels are zeroed when energy exceeds the threshold. Long waveforms support multi-crop evaluation, where each crop is processed independently and results are averaged. The SC model serves as a frozen teacher providing class guidance to TSE, supporting three label modes: the default "soft" mode passes SoftMax probability distributions, "one hot" passes hard labels, and "embedding" passes raw embedding vectors, ensuring that TSE benefits from SC's discrimination ability while remaining robust to prediction errors.

3. EXPERIMENTAL SETUP

3.1. Loss functions

The loss functions and weighting methods utilized in the training of each module are described below. The USS (Universal Source Separation) model's training objective is a multi-term weighted loss function, consisting of base separation/classification losses and optional semantic-acoustic bridge losses, all governed by configurable lambda weights that can be dynamically scheduled at runtime. The overall loss can be written as:

$$L_{foreground} + 0.01 \cdot \left(L_{interference} + L_{noise}\right) + 0.1 \cdot L_{cls} + L_{bridge} + L_{cross} + other\ auxiliary\ terms \quad (1)$$

The core foreground separation loss employs a Class-Aware Permutation Invariant Training (CA-PIT) assignment strategy, which jointly considers pairwise scale-invariant SDR improvement (SDRi) between estimated and reference foreground waveforms and pairwise negative log-likelihood of the predicted class logits at the reference class index, finding the optimal slot-to-reference permutation that minimizes their weighted sum. A confidence threshold filters out low-confidence class predictions from the valid pairing mask, and invalid class pairs incur a penalty cost. The matched foreground waveform loss (SA-SDRi)[12]is the primary separation objective, supplemented by a class cross-entropy term computed from the matched class NLL, and an inactive foreground energy penalty that drives energy of unmatched prediction slots toward zero. For silent segments, outlier exposure is applied to encourage a uniform class distribution using Kullback-Leibler (KL) divergence loss[13]. The silence decision is trained as a binary classification task using binary cross-entropy (BCE) loss. Interference and noise stems each have their own SI-SDR waveform losses with PIT assignment, inactive energy penalties, and optional temporal activity losses, weighted by λ_non_foreground=1.3 relative to the foreground loss. Additional auxiliary losses include a foreground count prediction loss, a DoA cosine similarity loss, a spatial diversity loss that penalizes same-class foreground slots whose spatial embeddings are too similar, and a waveform anti-collapse loss that penalizes high waveform correlation between same-class foreground slots. The semantic-acoustic bridge adds five further terms: a prototype loss, a supervised contrastive loss on aligned foreground embeddings, a bidirectional InfoNCE loss between audio and semantic embeddings, a bridge DoA alignment loss, and an embedding norm regularizer encouraging unit-norm embeddings. Two cross-contamination penalties are also applied: a cross-talk loss that penalizes cosine similarity between different-class foreground slot waveforms, and a cross-stem leakage loss that penalizes correlation between each stem's estimate and the wrong reference stem (e.g., foreground estimate correlating with interference reference).

The TSE loss is fundamentally a waveform-level SNR loss that ensures accurate reconstruction of extracted target sources.

3.2. DCASE2026 Task 4 Dataset

The development set is divided into training, validation, and test splits. The training split provides isolated source recordings, room impulse responses (RIRs), background noises, and interference sounds; mixtures are synthesized on-the-fly using SpAudSyn. The validation split (1,800 mixtures) is distributed with fixed metadata for deterministic reconstruction. The test split (1,512 mixtures) consists of fixed pre-synthesized mixtures. In both validation and test splits, 16.7% of mixtures contain no target, 16.7% contain one target, 33.3% contain two targets, and 33.3% contain three targets; within the two and three-target subsets, 50% contain duplicate same-class sources.

Each mixture is a 10-second, 4-channel first-order Ambisonics signal at 32 kHz, containing 0–3 target sound events (SNR 5–20 dB), 0–2 interference events (SNR 0–15 dB), and optional background noise, with at most 3 overlapping events. When same-class targets co-occur, their directions are separated by at least 60°. The dataset implementation (DatasetS3) supports three modes: "generate" for on-the-fly training synthesis with a configurable DUPE mechanism that duplicates same-class events at spatially separated positions (probability 0.5, minimum angular separation 60°), "metadata" for deterministic validation reconstruction, and "waveform" for direct loading of pre-synthesized mixtures at inference. Each sample provides a fixed number of source slots (default 3), padded with silence for inactive positions, with labels encoded as a stacked [n_sources, n_classes] matrix.

3.3. External Dataset

In addition to the provided development set, we also utilized data from the AudioSet strong portion of the AudioSet dataset[14].

4. EXPERIMENTAL RESULTS

Table 1 presents an overview of the four systems we submitted to the challenge. System S1 combines the outputs of the USS model and SC model into TSE model, and uses the outputs of TSE model as the final separated waveforms, it uses the SC classification result as the class labels. System S2 is basically the same as System S1, with the SC energy threshold stricter to filter less silence slots than System S1. System S3 uses the USS outputs as the separated waveforms, and uses SC classification result as the labels. System S4 uses exclusive mask for reducing the leakage of slots. Among the four systems, S1 and S3 achieve the highest Label Accuracy,

Table 1: Experimental results of the proposed framework

ID	Label Acc ↑	CAPI-SDRi ↑
S1	70.136	11.728
S2	70.068	11.743
S3	70.136	11.739
S4	65.013	11.262

while S2 attains the best CAPI-SDRi (11.743 dB). The performance gap between S1 and S3 is marginal in both metrics, suggesting that the TSE refinement provides modest gains on the validation set. System S4 exhibits notably lower performance in both metrics, indicating that the exclusive masking strategy, while reducing leakage, may overly constrain the mask flexibility and degrade separation quality.

5. CONCLUSION

This technical report presented our system for DCASE2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). The proposed system addresses two key challenges introduced in this year's task—same-class source ambiguity and zero-target mixtures—through a three-stage cascade pipeline consisting of universal sound separation (USS), audio tagging (AT), and iterative target sound extraction (TSE).

The core separation architecture employs TF-Locoformer, which enhances the standard Transformer by replacing linear layers in the feed-forward network with 1D convolutional and deconvolutional layers for local modeling, while self-attention captures global patterns. Placing such convolutional FFNs before and after self-attention in architecture enables the model to effectively combine both local and global information in the time-frequency domain. For universal sound separation, class-aware permutation invariant training (CA-PIT) is adopted to handle the ambiguity of same-class sources, while outlier exposure with KL divergence loss encourages uniform class distributions for silent segments. The silence decision is formulated as a binary classification task.

The TSE module leverages dual conditioning signals—enrollment audio for low-level acoustic matching and class clues via FiLM for high-level semantic guidance—enabling precise target extraction in complex acoustic environments. The iterative refinement strategy, where each iteration's output serves as enrollment for the next, progressively improves extraction quality. The M2D-SC classifier fuses a fine-tuned M2D backbone with three frozen pretrained SED encoders (BEATs, ATST-F, fPaSST) through a learnable fusion module, providing robust class guidance as a frozen teacher during TSE training.

Spatial features including IPD and ILD derived from the 4-channel input provide directional cues that enhance both separation and extraction performance. Future work may explore end-to-end joint training of all three stages, more sophisticated spatial feature extraction, and scaling the model to handle a larger number of overlapping sources.

6. REFERENCES

[1] Yasuda M, Nguyen BT, Harada N, Serizel R, Mishra M, Delcroix M, Hernandez-Olivan C, Araki S, Takeuchi D, Nakatani T, Ono N. Description and Discussion on DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes. arXiv preprint arXiv:2604.00776. 2026 Apr 1.

[2] Saijo K, Wichern G, Germain FG, Pan Z, Le Roux J. TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement. In2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC) 2024 Sep 9 (pp. 205-209). IEEE.

[3] Nguyen BT, Yasuda M, Takeuchi D, Niizumi D, Harada N. Class-Aware Permutation-Invariant Signal-to-Distortion Ratio for Semantic Segmentation of Sound Scene with Same-Class Sources. arXiv preprint arXiv:2601.22504. 2026 Jan 30.

[4] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in ICASSP 2023-2023 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), 2023, pp. 1–5.

[5] Schmid F, Morocutti T, Foscarin F, Schlüter J, Primus P, Widmer G. Effective pre-training of audio transformers for sound event detection. InICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025 Apr 6 (pp. 1-5). IEEE.

[6] Perez E, Strub F, De Vries H, Dumoulin V, Courville A. Film: Visual reasoning with a general conditioning layer. InProceedings of the AAAI conference on artificial intelligence 2018 Apr 29 (Vol. 32, No. 1).

[7] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, Pang R. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100. 2020 May 16.

[8] Lu Y, Li Z, He D, Sun Z, Dong B, Qin T, Wang L, Liu TY. Understanding and improving transformer from a multi-particle dynamic system point of view. arXiv preprint arXiv:1906.02762. 2019 Jun 6.

[9] Chen S, Wu Y, Wang C, Liu S, Tompkins D, Chen Z, Wei F. Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058. 2022 Dec 18.

[10] Li X, Shao N, Li X. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024 Jan 11;32:1336-51.

[11] Koutini K, Schlüter J, Eghbal-Zadeh H, Widmer G. Efficient training of audio transformers with patchout. arXiv preprint arXiv:2110.05069. 2021 Oct 11.

[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 Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2022, pp. 6022–6026.

[13] Kwon Y, Lee D, Kim D, Choi JW. Self-guided target sound extraction and classification through universal sound separation model and multiple clues. arXiv preprint arXiv:2509.13741. 2025 Sep 17.

[14] Bakhtin. Google's Audioset: Reformatted. Zenodo; 2022.