MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

An S5 System based on Multi-stage Separation and Classification

Saijo, Kohei; Masuyama, Yoshiki; Boeddeker, Christoph; Wichern, Gordon; Richter, Julius; Edo, Takahiro; Le Roux, Jonathan

TR2026-163 October 29, 2026

Abstract

This paper describes a multi-stage separation and classification system for DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). Inspired by the top-ranked system in DCASE 2025 Task 4, we adopt a cascaded framework consisting of universal sound separation (USS) with source counting, source classification, and class-aware refinement. In the first stage, a TF-Locoformer-based USS model separates multi-channel mixtures into single-channel foreground and interference signals. Each separated signal is then classified as one of 18 foreground classes or as interference. The separated foreground signals are further refined by another TF-Locoformer-based model conditioned on the predicted class labels and the observed mixture. Our best system achieves a CA-PI-SDRi of 12.94 dB on the final evaluation set, ranking third overall. Notably, the system achieves a label prediction accuracy of 76.92%, the best classification performance among all submissions.

Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2026

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

An S5 System based on Multi-stage Separation and Classification

Kohei Saijo[*1], *Yoshiki Masuyama*[*2], *Christoph Boeddeker*[2], *Gordon Wichern*[2], *Julius Richter*[2], *Takahiro Edo*[2], *Jonathan Le Roux*[2]

[1] Information Technology R&D Center, Mitsubishi Electric Corporation, Kanagawa, Japan,
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

***Abstract*—This paper describes a multi-stage separation and classification system for DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). Inspired by the top-ranked system in DCASE 2025 Task 4, we adopt a cascaded framework consisting of universal sound separation (USS) with source counting, source classification, and class-aware refinement. In the first stage, a TF-Locoformer-based USS model separates multi-channel mixtures into single-channel foreground and interference signals. Each separated signal is then classified as one of 18 foreground classes or as interference. The separated foreground signals are further refined by another TF-Locoformer-based model conditioned on the predicted class labels and the observed mixture. Our best system achieves a CA-PI-SDRi of 12.94 dB on the final evaluation set, ranking third overall. Notably, the system achieves a label prediction accuracy of 76.92%, the best classification performance among all submissions.**

***Index Terms*—S5, separation, counting, classification**

1. INTRODUCTION

The Spatial Semantic Segmentation of Sound Scenes (S5) task [1] aims to separate and classify target-class sound sources in multichannel mixtures. In DCASE 2026 Challenge Task 4 [2], the target classes comprise 18 predefined classes, and the numbers of target and interfering sources vary across mixtures. Since multiple target sources from the same class may be present, an S5 system must perform three processes: counting, separation, and classification.

Various approaches have been proposed for how and in what order these processes are performed. The challenge baseline first performs mixture-level tagging and then separates target sources based on the estimated classes [2], while introducing a "silence" class to jointly handle classification and counting. In the 2025 challenge [3], Morocutti et al. [4] proposed using frame-level sound event detection (SED) instead of tagging and achieved strong classification performance. However, this approach is not straightforward to apply to the 2026 challenge, where multiple sources from the same class may be present. Another line of work first separates the mixture and then classifies each separated signal; this approach won the 2025 challenge with strong separation performance [5], although its classification performance still leaves room for improvement.

Because many separation models can output only a fixed number of sources, separation-then-classification approaches require an additional source counting mechanism. A common strategy assumes a known maximum number of sources, trains the model to output near-zero signals for unused slots, and estimates the count using energy-based thresholding [6]. This strategy was also adopted in [5], however, it requires careful class-dependent threshold tuning and is sensitive to residual interference. These issues can be avoided by allowing the separation model itself to estimate the source count and output the corresponding number of sources [7], [8]. For example, [7] jointly trains a separator and a count estimator based on the separator's intermediate representations. It uses a source-count-dependent linear layer to produce the features to be decoded. Such a model enables the S5 task to be addressed by simply classifying each separated signal.

*Equal contribution.

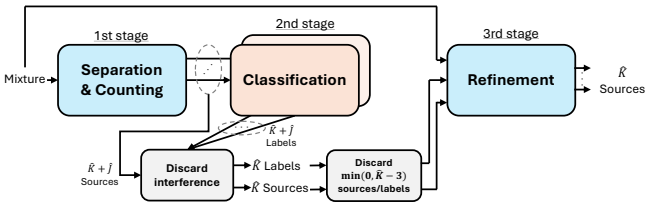

Fig. 1: Overview of our system.

Based on these considerations, we propose a multi-stage system for the S5 task, consisting of joint separation and counting, classification, and refinement. In the first stage, we employ a joint separation and counting model that extends a TF-Locoformer-based separation model [9], one of the state-of-the-art architectures, by incorporating a counting mechanism based on a classification (CLS) token. This model outputs separated signals corresponding to the estimated number of sources. The classification model then assigns a class label to each separated signal. Finally, we apply a refinement model that performs separation once again using the mixture, the separated signals, and the estimated class labels as inputs, producing the final separated signals.

2. METHOD

We consider an M-channel signal $\boldsymbol{y} \in \mathbb{R}^{M\times L}$ as a mixture of K target sources $\boldsymbol{s}_k$, J interfering sources $\boldsymbol{v}_j$, and a noise $\boldsymbol{b}$:

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{s}_k + \sum_{j=1}^{J} \boldsymbol{v}_j + \boldsymbol{b}, \tag{1}$$

where $K \in \{0,1,2,3\}$, $J \in \{0,1,2\}$, and empty sums are defined as zero: $\sum_{k=1}^{K} \boldsymbol{s}_k = 0$ if $K=0$ and $\sum_{j=1}^{J} \boldsymbol{v}_j = 0$ if $J=0$. All signals are represented in first-order Ambisonics (FOA) format (i.e., $M=4$) in DCASE 2026 Challenge Task 4. Target sources are defined as those belonging to one of the 18 predefined classes, while interfering sources are sound sources from all other classes. The goal is to separate K target sources on the W-channel, which represents the omnidirectional, direction-independent sound pressure component of the sound field, and to estimate the class label for each separated source.

2.1. First-stage model: Joint separation and counting

The first-stage model jointly performs separation and counting. Because it does not distinguish target from interfering sources, it separates all non-background sounds, including interfering sources, and estimates their total number, $N := K + J$. The separator is based on TF-Locoformer [9] with a Spectral Feature Compression (SFC) encoder and decoder [10], which we extend from fixed-source separation to variable-source separation with source-count estimation using a strategy similar to [7].

The encoder takes an M-channel short-time Fourier transform (STFT)-domain signal $\boldsymbol{Y} \in \mathbb{R}^{2M\times F\times T}$, where F and T denote the numbers of frequency bins and time frames, and the factor 2 represents the real and imaginary parts. The SFC encoder maps $\boldsymbol{Y}$ to

a D-dimensional feature $\boldsymbol{Z} \in \mathbb{R}^{D \times F' \times T}$ while reducing the frequency length to $F' (< F)$ via cross-attention with F' learnable queries.

Next, a learnable CLS token with a length of T' for source counting, $\boldsymbol{C} \in \mathbb{R}^{D \times 1 \times T'}$, is concatenated with the encoded features along the time dimension, with broadcasting along the frequency dimension. The resulting sequence is processed by $B_1^{(1\mathrm{st})}$ TF-Locoformer blocks:

$$[\boldsymbol{C}, \boldsymbol{Z}] \leftarrow \texttt{TFLocoformer_Blocks}_1([\boldsymbol{C}, \boldsymbol{Z}]). \tag{2}$$

Then, $\boldsymbol{C} \in \mathbb{R}^{D \times F' \times T'}$ is fed into a source-count estimator. The estimator first applies multiple repetitions of a Conv1D-based block followed by pooling along the frequency dimension to transform $\boldsymbol{C}$ into a $D' \times T'$ feature. It then averages the resulting feature over T', and predicts the total number of target and interfering sources using a linear layer followed by a softmax over six classes, corresponding to 0 to 5 sources. Subsequently, a linear layer is applied to $\boldsymbol{Z}$ to estimate as many feature representations as estimated source count $\hat{N}$:

$$\boldsymbol{Z} \leftarrow \texttt{Linear}_{\hat{N}}(\boldsymbol{Z}),\ \texttt{Linear}_{\hat{N}} : \mathbb{R}^{D \times F' \times T} \mapsto \mathbb{R}^{\hat{N}D \times F' \times T}. \tag{3}$$

Here, the model has five linear layers corresponding to $\hat{N} \in \{1, \ldots, 5\}$ with output size of $\hat{N}D$. By reshaping the output to $\boldsymbol{Z} \in \mathbb{R}^{\hat{N} \times D \times F' \times T}$, we obtain features $\boldsymbol{Z}_n \in \mathbb{R}^{D \times F' \times T}$ for each source with $n \in \{1, \ldots, \hat{N}\}$. Note that this linear layer and the subsequent blocks are not applied to samples for which $\hat{N} = 0$.

The $\hat{N}$ features are further processed by $B_2^{(1\mathrm{st})}$ TF-Locoformer blocks, where the weights of these blocks are shared across all sources, and the blocks are applied independently to each feature $\boldsymbol{Z}_n$:

$$\boldsymbol{Z}_n \leftarrow \texttt{TFLocoformer_Blocks}_2(\boldsymbol{Z}_n). \tag{4}$$

Finally, the SFC decoder receives $\boldsymbol{Z}_n$ and estimates separated sources $\hat{\boldsymbol{S}}_n \in \mathbb{R}^{2 \times F \times T}$, where frequency dimension is uncompressed by cross-attention with F learnable queries.

The model is trained using the sum of the time-domain thresholded SNR loss [11] on the separated signals, $\mathcal{L}_{\mathrm{snr}}$, and the cross-entropy loss for counting, $\mathcal{L}_{\mathrm{ce}}$, i.e., $\mathcal{L}_{\mathrm{snr}} + \gamma \mathcal{L}_{\mathrm{ce}}$, where γ is a weighting term.

Our design follows [7] in using a counting estimator and a source-count-dependent linear layer to separate a variable number of sources. However, whereas [7] feeds the mixture representation to the estimator, which corresponds to $\boldsymbol{Z}$ in (2) in our formulation, we introduce CLS-token-based counting. This is motivated by preliminary experiments showing that, with $\boldsymbol{Z}$-based counting, increasing γ degraded separation performance, suggesting a trade-off caused by sharing $\boldsymbol{Z}$ for separation and counting. By decoupling these representations, the proposed CLS-token-based approach is expected to mitigate this trade-off.

2.2. Second-stage model: Classification on separated signals

At the second stage of our system, as illustrated in Fig. 1, we predict a single class label for each separated signal. The first-stage separator outputs all non-background sources without distinguishing between target and interfering sources. Hence, the classifier considers not only the 18 target classes but also an additional "Interference" class. The combination of the first two stages yields all three processes for S5: counting, separation, and classification.

We construct five classification models based on general-purpose audio representation models using `PretrainedSED` [12] and ensemble their predictions. Specifically, we use BEATs [13], M2D [14], and ASiT [15] with the checkpoints provided by `PretrainedSED`[1]. PE-A-Frame-small [16] and PE-A-Frame-base [16] are also adopted using the checkpoints available on Hugging Face[2]. Each audio representation model comprises a feature extractor (`Enc`) and Transformer blocks (`Transformer`):

$$\hat{\boldsymbol{E}}_n = \texttt{Enc}(\hat{\boldsymbol{s}}_n), \tag{5}$$

$$\hat{\boldsymbol{V}}_n = \texttt{Transformer}(\hat{\boldsymbol{E}}_n), \tag{6}$$

where $\hat{\boldsymbol{s}}_n \in \mathbb{R}^L$ denotes the n-th separated signal on the W-channel. For M2D, the feature extractor computes log-mel spectrograms, while BEATs and ASiT rely on similar filter-bank features. PE-A-Frame instead uses DAC-VAE [17], which encodes time-domain signals into a continuous latent space. Then, the classification head averages the feature over time frames and predicts a soft label for the n-th source with softmax $\hat{\boldsymbol{o}}_n \in \mathbb{R}_+^{19}$.

The pretrained Transformer blocks and the classification head are jointly optimized to minimize the cross-entropy loss. Except for PE-A-Frame, we apply mixup augmentation in the feature domain. Specifically, we linearly interpolate two extracted features and the corresponding target one-hot vectors:

$$\hat{\boldsymbol{E}}_{\mathrm{mix}} = \alpha \hat{\boldsymbol{E}}_{n_1} + (1 - \alpha) \hat{\boldsymbol{E}}_{n_2}, \tag{7}$$

$$\boldsymbol{o}^{\star}_{\mathrm{mix}} = \alpha \boldsymbol{o}^{\star}_{n_1} + (1 - \alpha) \boldsymbol{o}^{\star}_{n_2}, \tag{8}$$

where $\alpha \sim \mathrm{Beta}(0.2, 0.2)$ is a mixing coefficient, and $\boldsymbol{o}^{\star}_n$ denotes the true one-hot vector for the n-th source signal.

2.3. Third-stage model: Refinement

Although the S5 task can be addressed using only the first- and second-stage models, we further apply a refinement model to improve separation quality, inspired by [5], [18]. The refinement model has an architecture similar to that of the first-stage separation model, but additionally uses the first-stage separated signals and the second-stage class labels as auxiliary inputs.

Specifically, the SFC encoder input is constructed by concatenating the first-stage separated signals with the multichannel mixture, and the resulting signals are jointly encoded into D-dimensional features. Following [18], we use signals refined by a multi-channel multi-frame Wiener filter (MCMFWF), rather than the raw first-stage outputs.

We modify the separator into a prompt-driven model [19]. Instead of a CLS token, the model uses class-specific learnable prompt vectors of size $D \times 1 \times 1$. Based on the class labels estimated in the second stage, the corresponding $\hat{K}$ prompts are concatenated with $\boldsymbol{Z}$ after broadcasting along the frequency dimension. After $B_1^{(3\mathrm{rd})}$ TF-Locoformer blocks, $\hat{K}$ features are obtained by taking the element-wise product between the transformed k-th prompt $\boldsymbol{P}_k \in \mathbb{R}^{D \times F' \times 1}$ and $\boldsymbol{Z} \in \mathbb{R}^{D \times F' \times T}$, broadcasting along the time dimension. This replaces the source-count-dependent linear layer in (3). The remaining processing is identical to the first-stage model: $B_2^{(3\mathrm{rd})}$ TF-Locoformer blocks and the SFC decoder are applied to each source.

We consider two variants for refinement. The multi-input multi-output (MIMO) system estimates all target sources simultaneously using all estimated sources and class labels, whereas the single-input single-output (SISO) system extracts each target source separately using only its corresponding estimated signal and class label together with the mixture. The MIMO system can exploit information from all target sources, while the SISO system is expected to be more robust to errors in the auxiliary separated signals and class labels.

3. EXPERIMENTS

3.1. Experimental setup for separation models

For the first- and third-stage models, we set $(B_1^{(1\mathrm{st})}, B_2^{(1\mathrm{st})}) = (8, 4)$ and $(B_1^{(3\mathrm{rd})}, B_2^{(3\mathrm{rd})}) = (6, 3)$, respectively. In both models, the first B_1 TF-Locoformer blocks used $D = 64$, $C = 384$, $K = 4$, $S = 1$,

[1] https://github.com/fschmid56/PretrainedSED

[2] https://huggingface.co/facebook/pe-a-frame-base

Table 1: Evaluation results of four systems. CA-PI-SDRi is reported in dB, and all other metrics are reported in %.

	Validation				Dev-test			Evaluation	
System	Count. acc.	CA-PI-SDRi	Acc. (mix)	Acc. (src)	CA-PI-SDRi	Acc (mix)	Acc (src)	CA-PI-SDRi	Acc. (mix)
Baseline [2]	-	8.26	56.56	67.99	8.49	60.71	70.39	6.76	55.16
System 1	**92.17**	14.25	72.61	77.36	14.84	78.11	83.52	12.77	75.00
System 2	**92.17**	14.36	73.22	**77.69**	14.77	**79.10**	**84.08**	12.77	**76.92**
System 3	**92.17**	**14.55**	73.22	**77.69**	**14.95**	**79.10**	**84.08**	**12.94**	**76.92**
System 4	91.83	14.37	**73.50**	76.43	14.41	76.06	81.63	12.54	73.28

$H = 4$, $G = 8$, and attention dimension $E = 128$[3]. In the subsequent B_2 blocks, C and E were changed to 256 and 96, respectively, while the other parameters were kept unchanged. We used the no-positional-encoding (NoPE) variant of the TF-Locoformer block [20] for the first-stage model and the original RoPE-based block [21] for the third-stage model. The STFT window and hop sizes were 1024 and 512 samples, respectively, and F' was set to 48.

The separation models were trained in two stages: unsupervised pretraining followed by supervised fine-tuning. We first pretrained the first-stage model with $(B_1^{(1\mathrm{st})}, B_2^{(1\mathrm{st})}) = (6, 3)$ on AudioSet [22] using mixture invariant training (MixIT) [11]. Because the number of sources in each mixture was unknown, we fixed the estimated number of sources to $\hat{N} = 8$, i.e., the linear layer in Eq. (3) produces outputs corresponding to eight sources. Pretraining was performed using monaural inputs, with a batch size of 32, 6 s inputs, and 262.5 k training steps. AdamW [23] with a fixed learning rate of 10^{-3} and a weight decay of 10^{-2} was used.

The first- and third-stage models were then initialized from the pretrained model and fine-tuned on the DCASE 2026 Task 4 dataset [2]. Parameters absent from pretraining, such as the CLS token in the first-stage model and the class-specific learnable prompts in the third-stage model, were randomly initialized. Because the first-stage model has more blocks than the pretrained model, the additional blocks were initialized by copying the final pretrained block of each stage: the sixth B_1 block for the seventh and eighth B_1 blocks, and the third B_2 block for the fourth B_2 block. Fine-tuning used 10 s inputs, a batch size of 32, and 150 k steps. The optimizer settings followed pretraining, except that a cosine learning-rate decay was used. Teacher forcing was applied by providing the ground-truth source counts and class labels during training. The loss weight γ was set to 1.

During inference, source-count estimation in the first stage was performed using either a single model or an ensemble of ten models trained with slightly different configurations, such as different batch size and CLS-token lengths $T' \in \{1, 2, 4, 8\}$. For the ensemble, we averaged the output probabilities of their source-count estimators.

3.2. Experimental setup for classification model

The classification models operate on the first-stage separated signals and thus must be robust to residual interference and separation artifacts. To improve robustness, we train the classification models using not only ground-truth sources (the W-channel of $\boldsymbol{s}_k$ and $\boldsymbol{v}_j$) but also separated signals $\hat{\boldsymbol{s}}_n$ from the first-stage models. In particular, the separation results from two models are used for training, where the two separators differ from those used in our final systems due to time limitation during the challenge. Exposing the classifiers to outputs produced by multiple separators should improve the robustness against the artifacts and the mismatch of the separators.

Since the number of isolated source signals in the official dataset is limited, we augment the training set using external resources. We retain up to 1000 clips with a single label per target class from AudioSet. We additionally collect up to 100 single-label clips per interference class. We further incorporate single-source recordings from NIGENS [24], SINS [25], and STARSS23 [26]. Specifically, we extract the segments with only "VacuumCleaner" from SINS, and "Clapping", "Footsteps", or "Pouring" from STARSS23 based on the official metadata. NIGENS provides the isolated signals for "AlarmClock" and "FootSteps". All signals are resampled to the sampling rate used in the pretraining of the audio representation models, i.e., 16 kHz for BEATs, M2D, and ASiT, and 48 kHz for PE-A-Frame. In addition to mixup, we also apply frequency warping and filter augmentation to the features, following the implementations in `PretrainedSED` [12].

We fine-tune each model using AdamW [23] with 1000 warm-up steps. The peak learning rates are set to 4.0×10^{-4} for the classification head and 4.0×10^{-5} for the Transformer blocks. We apply exponential moving average (EMA) to the models with a decay factor of 0.999 and select the EMA checkpoint with the best F1 score mean over classes for inference. For the ensemble, we take the average across the five models' outputs after taking softmax.

[3]The notation follows the original TF-Locoformer paper [9] and is independent of that used elsewhere in this paper.

Table 2: Ablation study of first-stage model on validation set. Pretrain indicates whether model is pretrained with MixIT. BEATs is used for classification.

Counting input	Pretrain	Count. acc.	SDRi	CA-PI-SDRi
Feature $\boldsymbol{Z}$	✓	91.39	16.95	13.50
CLS token $\boldsymbol{C}$ ($T' = 1$)	✓	91.17	17.08	13.70
CLS token $\boldsymbol{C}$ ($T' = 2$)	✓	91.10	17.07	13.71
CLS token $\boldsymbol{C}$ ($T' = 4$)	✓	**92.44**	**17.13**	**13.75**
CLS token $\boldsymbol{C}$ ($T' = 8$)	✓	91.44	17.12	13.73
CLS token $\boldsymbol{C}$ ($T' = 4$)		91.44	16.09	12.67

Table 3: Classification accuracy on dev-test set.

Model	Acc. (mix)	Acc. (src)
BEATs-clean	65.74	72.31
BEATs	76.32	81.78
M2D	75.60	81.54
ASiT	74.01	79.71
PE-A-Frame (small)	75.27	80.30
PE-A-Frame (base)	73.88	79.63
Ensemble (5 models)	**79.10**	**84.08**

3.3. Results

We evaluated the following four systems. All systems except System 4 used an ensemble of multiple model outputs for source counting and classification. When four or more sources were identified as target sounds, Systems 2–4 discarded sources before inputting them to the refinement model so that the number of sources was reduced to three. Specifically, the sources were discarded in ascending order of the maximum estimated class probability, $\arg\max(\hat{\boldsymbol{o}}_n)$.

System 1: A system that uses only the first- and second-stage models, without applying the third-stage refinement.

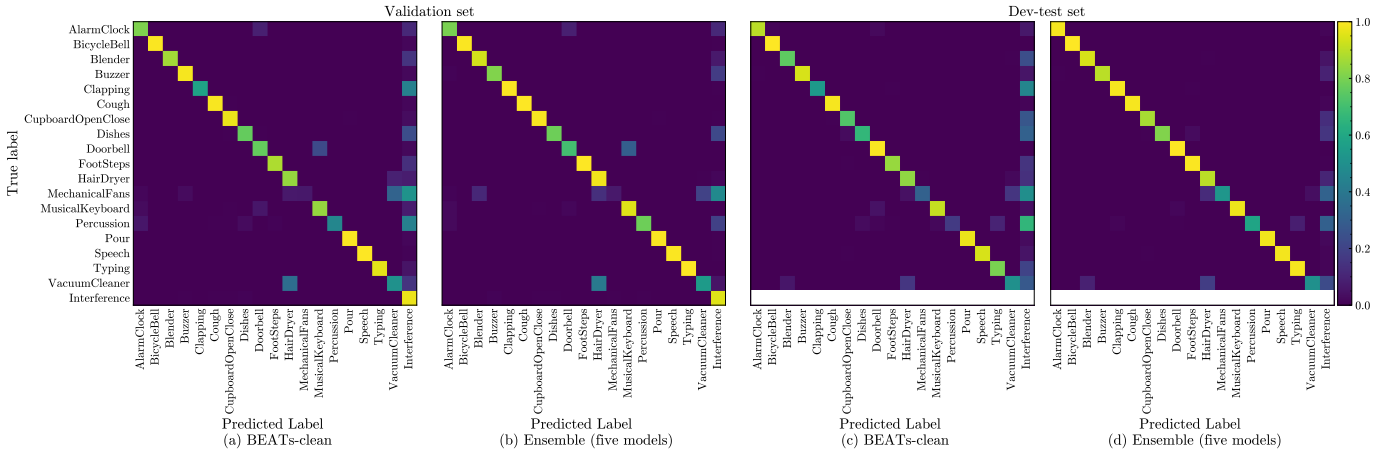

Fig. 2: Confusion matrices for BEATs-clean (the first row in Table 3) and ensemble of the five models (the last row in Table 3) with row-wise normalization. For this figure, no sources were discarded even when four or more sources were estimated as target sources.

System 2: The three-stage system described in Section 2, where the MIMO version is used for the third-stage refinement model.
System 3: A system in which the third stage of System 2 is replaced with the SISO version.
System 4: The same pipeline as System 3, but without ensembling the source-counting and classification results.

Table 1 reports the source-counting accuracy (Count. acc.), class-aware permutation-invariant SDR improvement (CA-PI-SDRi) [27], and classification accuracies at the mixture level (Acc. (mix)) and source level (Acc. (src)) [2]. Since ground-truth sources are available only for the validation set, counting accuracy is reported only for the validation set. Note that, although Systems 1–3 use exactly the same ensemble system for both source counting and classification, their accuracies differ because Systems 2 and 3 discard sources when four or more sources are identified as target sounds.

Table 1 shows that our system substantially outperforms the baseline [2] in both separation and classification. The first-stage model (System 1) already achieves reasonable counting accuracy and separation performance, and adding the refinement model yields modest improvements. This may be because the first-stage model already benefits from unsupervised pretraining, while the refinement model has a larger mismatch between pretraining and fine-tuning due to structural differences, such as randomly initialized learnable prompts and the introduction of RoPE. Further analysis is left for future work. The results of Systems 3 and 4 indicate that ensembling improves both counting and classification.

Table 2 presents ablation results for the first-stage model. All experiments used a model with $B_1^{(1\mathrm{st})} = 6$ and $B_2^{(1\mathrm{st})} = 3$, with the batch size reduced to 16. SDRi was computed using the optimal permutation between the estimated and ground-truth target and interfering sources. Unlike CA-PI-SDR, this metric does not penalize class-estimation errors, but it penalizes the counting errors in the same penalties as CA-PI-SDR. To evaluate the proposed CLS-token-based source-count estimation, we compared it with a baseline that estimates the source count from the feature representation $\boldsymbol{Z}$, following prior work [7]. This baseline applies several Conv2D blocks to $\boldsymbol{Z}$ and performs temporal pooling before count estimation. As expected from the discussion in Section 2.1, the results show that the CLS-token-based approach improves separation performance while maintaining comparable or higher counting accuracy. The CLS-token length T' had only a minor effect, with $T' = 4$ performing best. For this model, MixIT pretraining substantially improved separation performance without noticeably affecting counting accuracy, consistent with prior findings on the effectiveness of unsupervised pretraining under limited supervised data [28], [29].

Table 3 shows classification accuracy for different second-stage models. Here, BEATs-clean denotes a BEATs-based classifier trained exclusively on ground-truth source signals. Comparison between the two BEATs-based systems validates the effectiveness of augmenting the training data with separated signals and additional isolated recordings. Among the individual audio representation models, BEATs achieves the best performance, while the ensemble of the five models yields a further improvement.

Figure 2 depicts the confusion matrices for BEATs-clean and the ensemble of five models, where rows and columns show the true and predicted labels, respectively. Here, we assign the true label for each separation output according to SNR with respect to the ground-truth target and interference sources. Since ground-truth interference sources are unavailable for the dev-test set, separation outputs that do not match any target source are excluded from the evaluation. Consequently, the last rows of panels (c) and (d) are empty. When the number of sources in a mixture is overestimated, unmatched separation outputs are also omitted from this visualization. The confusion matrices reveal clear diagonal patterns, but some target classes such as “MechanicalFans” and “Percussion” are misclassified as interference. These misclassifications are partially mitigated in the ensemble of the five models compared with BEATs-clean.

4. CONCLUSION

We presented a multi-stage separation and classification system for the S5 task, consisting of a TF-Locoformer-based first-stage model for joint separation and counting, a second-stage classifier, and a third-stage refinement model. Experiments showed that the proposed system substantially outperformed the official baseline in both separation and classification. Our best system achieved a CA-PI-SDRi of 12.94 dB on the final evaluation set, ranked third overall in DCASE 2026 Challenge Task 4, and achieved the highest classification accuracy among all submissions. Ablation results confirmed the effectiveness of CLS-token-based counting, MixIT pretraining, and ensembling. Although the refinement stage yielded only modest gains, cascading separation and counting, classification, and class-aware refinement remains a promising approach for S5. Future work will explore tighter integration of separation and classification and more robust refinement.

REFERENCES

[1] 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," in *Proc, EUSIPCO*, 2025, pp. 266–270.

[2] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, C. Hernandez-Olivan, S. Araki, D. Takeuchi, T. Nakatani *et al.*, "Description and discussion on DCASE 2026 challenge task 4: Spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2604.00776*, 2026.

[3] 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.

[4] 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., June 2025.

[5] 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.

[6] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *Proc. ICASSP*, 2021, pp. 186–190.

[7] J. Zhu, R. Yeh, and M. Hasegawa-Johnson, "Multi-decoder DPRNN: High accuracy source counting and separation," *arXiv preprint arXiv:2011.12022*, 2020.

[8] S. R. Chetupalli and E. A. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.

[9] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. Le Roux, "TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement," in *Proc. IWAENC*, 2024.

[10] K. Saijo and Y. Bando, "Input-adaptive spectral feature compression by sequence modeling for source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, 2026.

[11] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Proc. NeurIPS*, vol. 33, pp. 3846–3857, 2020.

[12] F. Schmid, T. Morocutti, F. Foscarin, J. Schlüter, P. Primus, and G. Widmer, "Effective pre-training of audio transformers for sound event detection," in *Proc. ICASSP*, 2025.

[13] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023, pp. 5178–5193.

[14] D. Niizumi, D. Takeuchi, M. Yasuda, B. Thien Nguyen, Y. Ohishi, and N. Harada, "M2D-CLAP: Exploring general-purpose audio-language representations beyond CLAP," *IEEE Access*, vol. 13, pp. 163 313–163 330, 2025.

[15] S. A. A. Ahmed, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "ASiT: Local-global audio spectrogram vision transformer for event classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3684–3693, 2024.

[16] A. Vyas, H.-J. Chang, C.-F. Yang, P.-Y. Huang, L. Gao, J. Richter, S. Chen, M. Le, P. Dollár, C. Feichtenhofer *et al.*, "Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning," in *Proc. CVPR*, 2026, pp. 30 172–30 182.

[17] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang *et al.*, "Movie Gen: A cast of media foundation models," *arXiv preprint arXiv:2410.13720*, 2024.

[18] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[19] K. Saijo, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Task-aware unified source separation," in *Proc. ICASSP*, 2025.

[20] K. Saijo and T. Ogawa, "A comparative study on positional encoding for time-frequency domain dual-path transformer-based source separation models," *Proc, EUSIPCO*, 2025.

[21] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[22] 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 *Proc. ICASSP*, 2017, pp. 776–780.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2018.

[24] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS general sound events database," *arXiv:1902.08314*, 2019.

[25] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, T. Van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. DCASE*, 2017.

[26] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. NeurIPS*, vol. 36, 2023, pp. 72 931–72 957.

[27] 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," *arXiv preprint arXiv:2601.22504*, 2026.

[28] A. Sivaraman, S. Wisdom, H. Erdogan, and J. R. Hershey, "Adapting speech separation to real-world meetings using mixture invariant training," in *Proc. ICASSP*, 2022, pp. 686–690.

[29] K. Saijo and Y. Bando, "Is MixIT really unsuitable for correlated sources? exploring MixIT for unsupervised pre-training in music source separation," in *Proc. WASPAA*, 2025, pp. 1–5.