A LABEL-GUIDED ENSEMBLE SYSTEM FOR SPATIAL SEMANTIC SEGMENTATION OF SAME-CLASS SOUND SOURCES

Technical Report

Yongyi Deng[1,†], *Tong Zou*[1,†], *Yanxin Tian*[1], *Hao Shi*[2], *Yicheng Yan*[1], *Jiayue Luo*[1], *Gongping Huang*[1*]

[1] School of Electronic Information, Wuhan University, Wuhan, China
[2] Kyoto University, Kyoto, Japan
[†] Equal contribution.

ABSTRACT

This report presents our system for DCASE 2026 Task 4, which addresses spatial semantic segmentation of sound scenes containing same-class foreground sources and inactive source labels. The task requires not only separating active sound events that may share the same class label, but also handling source slots corresponding to inactive labels. To improve the reliability of label-conditioned source separation, we adopt a label-guided ensemble strategy. In the tagging stage, two M2D-AT variants and one CLAPAT variant are fused by weighted voting to obtain robust source-label estimates. The estimated labels are then used to guide source separation. The primary separator is a fine-tuned ResUNetK model with a mask-sharpen inference variant, while a TF-GridNet model is used only as a weak auxiliary branch for a small number of selected classes through class-dependent fusion weights. Instead of uniformly averaging separator outputs, the final outputs are generated through label-guided class-dependent fusion, which improves the consistency between predicted labels and separated sources while keeping inactive slots controlled by silence-label conditioning. On the development set, according to the submission-pack evaluation snapshot, our final system achieves a CAPI-SDRi of 8.625, with a mixture-level accuracy of 61.706% and a source-level accuracy of 72.139%.

***Index Terms*—** label-guided source separation, audio tagging ensemble, class-conditioned separation

1. INTRODUCTION

DCASE 2026 Task 4 focuses on spatial semantic segmentation of sound scenes, where the system is required to separate foreground sound sources and assign semantic labels to the separated [1]. Compared with conventional source separation tasks, this task is more challenging because multiple foreground sources may belong to the same semantic class, while some source slots may correspond to inactive or silent labels. Therefore, the system should not only improve the signal quality of separated sources, but also maintain the consistency between separated waveforms and predicted semantic labels.

The official baseline follows a two-stage framework consisting of audio tagging and label-conditioned source separation [2]. In this framework, the tagging model first predicts the source labels, and the separation model then generates source estimates conditioned on the predicted labels. However, errors in the tagging stage can easily propagate to the separation stage. False positive labels may produce unreliable estimates for inactive sources, while false negative labels may cause active sources to be missed. In addition, different separation models may show different behavior across sound classes.

To improve the robustness of this framework, we propose a label-guided ensemble system. The system first fuses the predictions of multiple audio tagging models to obtain more reliable source-label estimates. These estimated labels are then used to guide a ResUNetK-dominated class-dependent fusion process, where a weak TF-GridNet auxiliary branch is only used for a small number of selected classes. The final system does not rely on an additional verification or relabeling stage; instead, inactive slots are mainly handled through the predicted silence label and zero label-vector conditioning.

2. PROPOSED SYSTEM

The proposed system consists of three main components: audio tagging ensemble, label-conditioned separation, and label-guided output fusion. The overall framework is shown in Fig. 1.

2.1. Audio Tagging Ensemble

The audio tagging module predicts the semantic label of each source slot. Since the following separation stage is conditioned on the predicted labels, robust label estimation is important for the final class-aware separation performance.

Our final system ensembles three audio tagging (AT) models: two M2D-AT variants and one CLAP-AT variant. The first M2D-AT variant adopts M2D as the audio encoder [3], with a BiGRU-based classification head for label prediction. The second M2D-AT variant follows the baseline architecture, using M2D as the audio encoder and the baseline's fully connected classification head. The third model is a CLAP-AT variant, which leverages CLAP's HTSAT audio encoder and adopts the same fully connected classification head as the baseline [4, 5]. By combining models with distinct audio encoders and classification head designs, the ensemble yields complementary predictions and enhances the robustness of sound source label estimation.

All three AT models follow a unified two-stage fine-tuning protocol. In the first stage, we freeze the full parameters of the pre-trained audio encoder and only optimize the classification head, enabling rapid adaptation of pre-trained audio representations to the sound source label prediction task. In the second stage, we unfreeze

*Corresponding author.

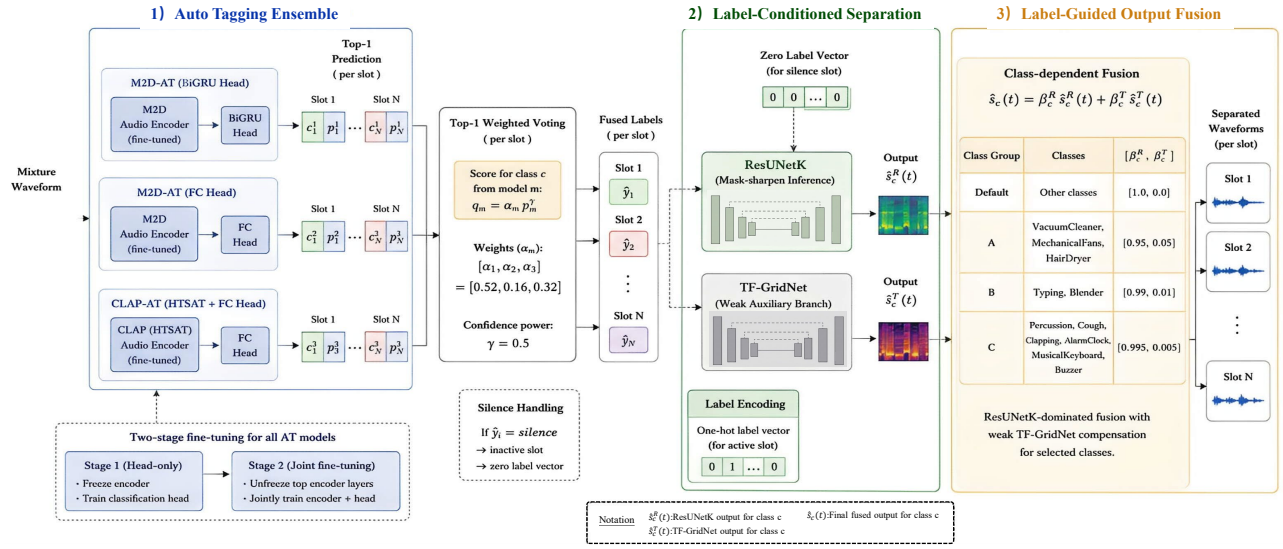

Figure 1: Overview of the proposed label-guided ensemble system.

the top layers of the audio encoder and conduct joint fine-tuning alongside the classification head, to further refine high-level acoustic feature extraction and improve overall tagging performance.

The predictions of the three models are fused by top-1 weighted voting. For each source slot, each tagging model provides its top-1 predicted label and the corresponding confidence score. If the m-th model predicts class c_m with confidence p_m, then this prediction contributes a score

$$q_m = \alpha_m p_m^{\gamma}, \tag{1}$$

to class c_m, where α_m is the model weight and γ is a confidence power. The final score of each candidate label is obtained by summing the contributions from all models that predict that label.

In our final system, the weights of the three tagging models are set to

$$[\alpha_1, \alpha_2, \alpha_3] = [0.52, 0.16, 0.32], \tag{2}$$

and the confidence power is set to $\gamma = 0.5$. The final label of each slot is selected as the class with the highest accumulated weighted score.

No additional global active-label threshold is used in the final configuration. Instead, silence is treated as a candidate label. If the fused top-1 label of a source slot is silence, the slot is regarded as inactive and is represented by silence-aware zero label-vector conditioning in the following separation stage.

2.2. Label-Conditioned Separation

The main separation backbone of our system is ResUNetK [6], implemented in the final submission with a mask-sharpen inference variant. We use a fine-tuned ResUNetK checkpoint as the primary separator. The final checkpoint is selected from a continued fine-tuning run, where the model is further trained from a previous fine-tuned ResUNetK checkpoint. In our final submission, the epoch-3 checkpoint of this continued run is used as the main separation model.

The final submission also includes a TF-GridNet separator as a weak auxiliary branch [7]. However, TF-GridNet is not used as the main separation model. In the final configuration, the default fusion weight is $[1.0, 0.0]$, which means that most classes rely entirely on the ResUNetK output. TF-GridNet is only assigned a very small weight for several selected classes. This design keeps the strong overall performance of ResUNetK while using TF-GridNet only as a weak class-dependent compensation branch rather than as a symmetric second main separator.

Given a predicted label c, the ResUNetK output and the optional TF-GridNet auxiliary output are denoted as

$$\hat{s}_c^{\mathrm{R}}(t), \quad \hat{s}_c^{\mathrm{T}}(t), \tag{3}$$

where $\hat{s}_c^{\mathrm{R}}(t)$ denotes the ResUNetK output and $\hat{s}_c^{\mathrm{T}}(t)$ denotes the TF-GridNet output.

2.3. Label-Guided Output Fusion

The final waveform is obtained by class-dependent weighted fusion:

$$\hat{s}_c(t) = \beta_c^{\mathrm{R}} \hat{s}_c^{\mathrm{R}}(t) + \beta_c^{\mathrm{T}} \hat{s}_c^{\mathrm{T}}(t), \tag{4}$$

where β_c^{R} and β_c^{T} are the fusion weights of ResUNetK and TF-GridNet, respectively.

For most classes, the weights are set to $[1.0, 0.0]$, so the final output is entirely taken from the ResUNetK branch. For a small number of classes, TF-GridNet is weakly fused with a very small weight. The final class-dependent fusion rules are summarized in Table 2.

These weights indicate that the final system is dominated by ResUNetK. The TF-GridNet branch is only used as a weak auxiliary component for class-dependent compensation. Compared with a strong ResUNetK-only configuration, this weak fusion improves CAPI-SDRi from 8.587 to 8.625 on the development set, while keeping the mixture-level and source-level accuracies unchanged.

For inactive slots, the silence label is mapped to a zero label vector, so inactivity is handled implicitly through silence-aware conditioning rather than by an additional verification module. The final system does not enable an additional verification or silence-gate post-processing stage. This makes the final pipeline simple and deterministic: the tagging ensemble determines the slot labels, and

Table 1: Development-set results of different system variants. Acc. Mix. and Acc. Src. denote mixture-level and source-level classification accuracy, respectively.

System	CAPI-SDRi	Acc. Mix.	Acc. Src.
Official ResUNetK baseline	8.286	57.474%	68.238%
Weighted label fusion without verification	8.317	61.045%	70.501%
Weighted label fusion with silence gate	8.379	61.574%	71.205%
CLAP-oriented silence-gate variant	8.419	61.971%	71.543%
Three-model tagging ensemble + fine-tuned ResUNetK	8.587	61.706%	72.139%
ResUNetK + weak class-dependent TF-GridNet compensation	8.616	61.706%	72.139%
Proposed final system	**8.625**	**61.706%**	**72.139%**

Table 2: Class-dependent weights for ResUNetK-dominated fusion with weak TF-GridNet compensation.

Group	Classes	Weights
Default	Other classes	$[1.0, 0.0]$
A	VacuumCleaner, MechanicalFans, HairDryer	$[0.95, 0.05]$
B	Typing, Blender	$[0.99, 0.01]$
C	Percussion, Cough, Clapping, AlarmClock, MusicalKeyboard, Buzzer	$[0.995, 0.005]$

the label-guided fusion module generates the corresponding separated waveforms.

3. EXPERIMENTS

3.1. Evaluation Metrics

We evaluate our system on the development set using the official evaluation script. The reported metrics include CAPI-SDRi, mixture-level accuracy, and source-level accuracy. CAPI-SDRi measures the class-aware separation quality, while the two accuracy metrics reflect the correctness of semantic label prediction. All model selection, fusion weights, and system variants are determined based on development-set experiments. No labels from the official evaluation set are used for model selection or parameter tuning.

3.2. Ablation Results

Table 1 summarizes the development-set performance of the main system variants. Starting from the official ResUNetK baseline, we first improve the label prediction stage by weighted label fusion. Several inactive-slot handling variants, including silence-gate and CLAP-oriented silence-gate strategies, are also evaluated. The final system keeps the three-model tagging ensemble and the fine-tuned ResUNetK separator, and further applies weak class-dependent TF-GridNet compensation for selected classes.

The results show that the largest improvement over the official ResUNetK baseline comes from more reliable label estimation and fine-tuning of the ResUNetK separator. The final weak TF-GridNet fusion further improves the CAPI-SDRi from 8.587 to 8.625, corresponding to an absolute gain of 0.038. Since the TF-GridNet weights are very small and only applied to selected classes, the final system should be regarded as a ResUNetK-dominated system with weak class-dependent auxiliary fusion, rather than a symmetric two-separator ensemble. The final score of 8.625 is taken from the final submission-pack evaluation snapshot and the corresponding rerun average-metrics file.

3.3. Final Submission

The final submitted system uses the three-model tagging ensemble described in Section 2.1 and the label-guided separation fusion described in Section 2.3. The main separation model is the fine-tuned ResUNetK checkpoint with the mask-sharpen inference variant, while the TF-GridNet branch is only used for selected classes with small fusion weights. According to the final submission-pack evaluation snapshot, the final system achieves a CAPI-SDRi of 8.625, a mixture-level accuracy of 61.706%, and a source-level accuracy of 72.139% on the development set.

The unchanged label accuracy between the fine-tuned ResUNetK system and the final system indicates that the final improvement mainly comes from separation-side compensation rather than changes in label prediction. This is consistent with the design of the weak TF-GridNet fusion, which only modifies the waveform estimates for selected classes while keeping the fused labels unchanged.

4. CONCLUSION

This report describes our submission system for DCASE 2026 Task 4. The proposed system follows a two-stage audio tagging and label-conditioned source separation framework. In the tagging stage, two M2D-AT variants and one CLAPAT variant are fused by top-1 weighted voting to obtain robust source-label estimates. In the separation stage, a fine-tuned ResUNetK model with a mask-sharpen inference variant is used as the primary separator, while a TF-GridNet model is used only as a weak auxiliary branch for selected classes through ResUNetK-dominated class-dependent fusion. The final system does not enable an additional verification or silence-gate post-processing stage. On the development set, according to the submission-pack evaluation snapshot, the proposed final system achieves a CAPI-SDRi of 8.625, with a mixture-level accuracy of 61.706% and a source-level accuracy of 72.139%.

5. REFERENCES

[1] 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

[2] 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 *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2391–2406, 2024.

[4] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[6] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data for computational auditory scene analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 206–219, 2026.

[7] 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 Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.