THE SYSTEM FOR DCASE 2025 SOUND EVENT LOCALIZATION AND DETECTION CHALLENGE

Technical Report

Chengnuo Sun

Jiangsu University
zhenjiang, China
2222408014@stmail.ujs.edu.cn

Lijian Gao

Jiangsu University
zhenjiang, China
ljgao@ujs.edu.cn

ABSTRACT

This technical report gives an overview of our system for task3 with audiovisual of the DCASE 2025 challenge. We propose a Sound Event Localization and Detection (SELD) system for stereo sound event localization and detection in regular video content. Compared with the baseline, the proposed method pays more attention to the temporal relationship between modalities. We evaluated our methods on the dev-test set of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset and we achieve significant improvements over the baseline method.

Index Terms— Sound Event Location and Detection, Sound Event Detection, Direction of Arrival Estimation, Mutlimodal

1. INTRODUCTION

Sound Event Localization and Detection (SELD) integrates temporal identification and classification of active acoustic sources (Sound Event Detection, SED) with spatial position or direction-of-arrival (DOA) estimation [1][2][3]. This multimodal capability underpins critical applications spanning human-robot collaboration, augmented reality systems, navigation platforms, intelligent home technologies, and security solutions. Research advancements have systematically addressed core challenges: detecting polyphonic auditory scenes [4], localizing concurrent same-class events involving mobile sound sources [5], and suppressing irrelevant external sounds [6]. Recent developments extend to active source distance estimation in current evaluations [7]. Although historically audio-centric, the DCASE challenge has introduced audio-visual tracks in its latest editions, leveraging the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [8, 9].

This inclusion enables the investigation of SELD as a multimodal audio-visual challenge. Vision and audio modalities offer complementary strengths: visual information delivers high spatial precision, whereas audio sensing can detect obscured sources. The present report details the systems we developed and submitted for the audio-visual track of this challenge. Specifically, we identify that existing fusion schemes (e.g., Transformer-based feature concatenation) fail to adequately model cross-modal temporal dependencies, while accurate Sound Event Localization and Detection (SELD) critically relies on capturing dynamic spatiotemporal evolution. To address this limitation, we propose an **Audio Temporal Enhancement Module** that explicitly captures complex temporal correlations within audio streams and their synchronization with visual sequences, thereby enriching multimodal fusion with enhanced contextual information. Furthermore, we demonstrate that directly employing generic visual features extracted from pre-trained ResNet-50 [10] introduces domain adaptation issues. These features frequently lack sufficient representation of spatial cues essential for sound event localization, such as object orientation and partial occlusion patterns. Consequently, we design a **Visual Feature Enhancement Module** that optimizes pre-extracted visual features to improve their representational fidelity and discriminative capacity for SELD-specific localization tasks.

Generally, the audio temporal enhancement module enables dynamic modal information interaction understanding during fusion, while the enhanced visual features provide task-oriented inputs. This framework substantially improves the robustness and localization accuracy of our system for audio-visual SELD.

2. METHOD

2.1. Audio Temporal Enhancement

The proposed Audio Temporal Enhancement Module incorporates an innovative multi-head attention architecture [11] to address the critical need for temporal modeling in audio feature representation. Specifically, the module employs a dual-head self-attention mechanism to perform parallelized temporal modeling of audio features [12]. This design enables simultaneous capture of feature dependencies across multiple temporal scales [13], effectively resolving the limitation of single-scale temporal modeling in conventional approaches [1].

In implementation, the module adopts a batch-first data organization scheme [14], optimizing memory alignment with modern deep learning frameworks to significantly accelerate computational efficiency. For feature interaction, the module utilizes a learnable query-key-value weighting mechanism [15] that dynamically assigns attention weights to temporal features. This adaptive weighting strategy allows the model to focus on salient temporal dependencies while suppressing noise components [16]. To regularize the model, a moderate dropout rate [17] is applied—a value empirically determined to balance representational capacity and overfitting prevention.

Crucially, the module forms an end-to-end spatiotemporal feature learning framework through synergistic integration with subsequent Transformer decoders [18]. This architectural coupling ensures:

1. Effective extraction of hierarchical temporal patterns in audio streams [19]

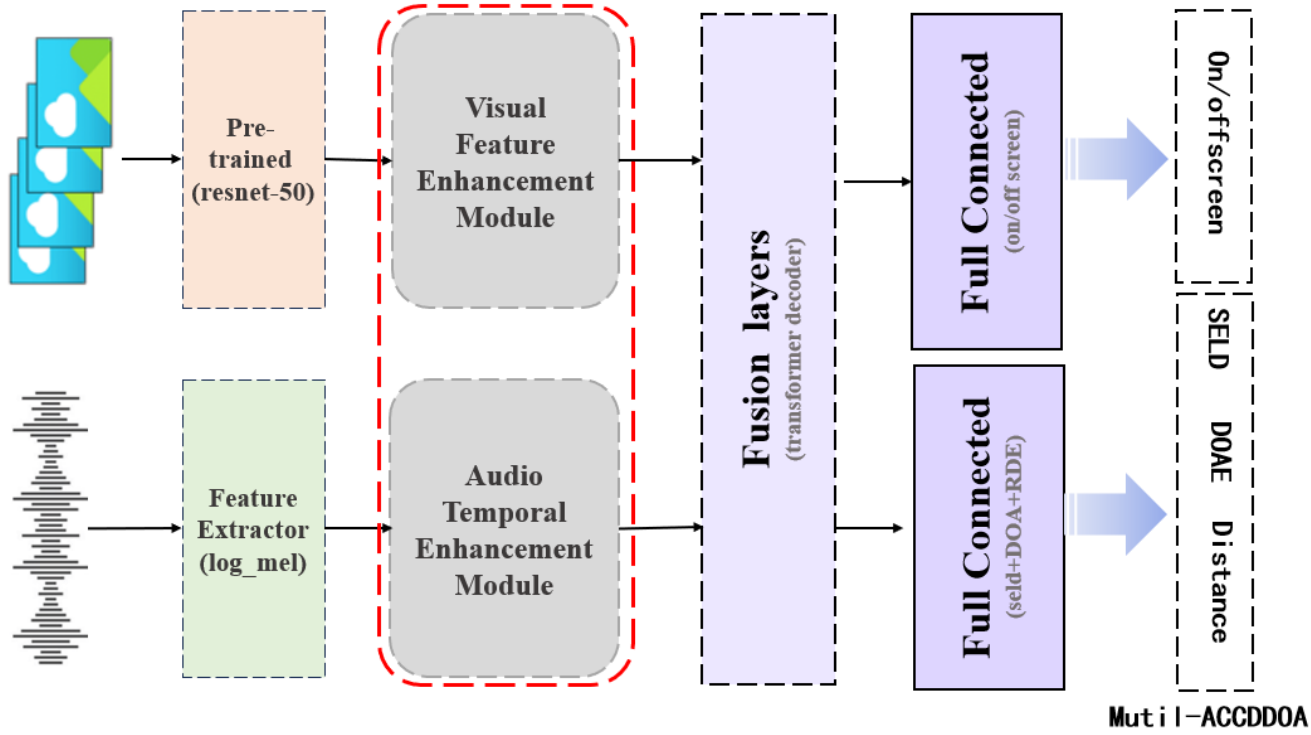

Figure 1: : A Overview of our SELD system.The framework include four parts: Feature Extraction, Enhancement Module, Fusion Layers and Full Connected Layers.

2. Generation of temporally coherent representations that serve as optimal inputs for cross-modal fusion layers

The design fundamentally bridges temporal feature extraction and multimodal integration within a unified computational graph [20].

2.2. Visual Features Enhancement

The proposed visual feature enhancement module is designed to address key challenges in multimodal representation learning [21]. The key resides in a systematically designed transformation pathway comprising three fundamental operations: First, a linear projection layer maps features extracted by the pretrained ResNet-50 model [10] into a unified latent space, achieving cross-modal alignment while preserving essential spatial topology [22]. Second, layer normalization (LayerNorm) [15] performs dimension-wise standardization to stabilize feature distributions across varying environmental conditions, effectively mitigating internal covariate shift during training [16]. Third, ReLU activation functions [23] introduce controlled non-linearity, enabling the modeling of complex visual pattern interactions. This processing cascade ensures consistent semantic representation across heterogeneous modalities while maintaining structural integrity of visual features.

The module further incorporates two specialized mechanisms to enhance representational robustness: Residual connections [10] orchestrate multi-resolution feature fusion, integrating shallow spatial details with deep semantic abstractions to form comprehensive hierarchical representations [24]. Complementarily, configurable dropout regularization [17] employs stochastic feature suppression to prevent overfitting while promoting feature diversity learning [25]. The resulting enhanced features exhibit optimized spatial-semantic coherence, providing superior inputs for downstream multimodal fusion modules [18]. This integrated design significantly advances joint audio-visual perception capabilities, particularly in complex spatial localization scenarios requiring precise environmental understanding [26].

2.3. System Design

This paper proposes an end-to-end Multimodal Sound Event Localization and Detection (SELD) framework [18], the core of which lies in a systematic multimodal co-processing architecture. The model consists of four modules that complement each other:

Audio Temporal Enhancement Module uses a cascading convolutional block (ConvBlock) for time-frequency domain feature learning [1], and each ConvBlock integrates two-dimensional convolution, batch normalization, ReLU activation, and maximum pooling operations in turn to effectively extract acoustic representations with spatiotemporal invariance. The module further models the temporal dynamic characteristics through the bidirectional GRU network, and introduces the multi-head self-attention mechanism [11] to capture the long-range context dependence.

Visual Feature Enhancement Module is designed with an innovative feature processing mechanism: cross-modal space alignment is achieved through linear projection, layer normalization operation [15] stabilizes feature distribution, and a configurable dropout mechanism is used to improve the generalization ability of the model. The design enhances feature discrimination while maintaining the visual semantic integrity.

Multimodal Fusion Module integrates the temporal-aware

Table 1: Performance comparison on STARSS23 dev-test split.

Method	$F_{20°/1/on}$ (%)	DOAE (°)	RDE (%)	OSA (%)
Official Baseline	20.0	23.8	40.0	80.0
Replicated Baseline	18.8	25.7	39.0	79.9
AV-SELD (ours)	**23.1**	**20.1**	**34.0**	79.1

Transformer decoder architecture [18] to realize the deep interaction of audio and video features through the cross-attention mechanism. The structure dynamically models cross-modal spatiotemporal associations to provide optimized feature representation for joint sensing.

Multi-task prediction head adopts a parallel output design: the DOA prediction branch constrains the three-dimensional spatial orientation output through the Tanh activation function; The distance estimation branch uses ReLU to ensure the non-negative characteristics of physical distance; The spatial positioning state branch uses Sigmoid to output the probability distribution of the target on and off the screen. Each branch uses a dedicated activation function to ensure the physical rationality of the output value.

The architecture supports flexible expansion of mono and multi-track scenarios, and its modular design shows significant generalization advantages and robust performance in complex acoustic environments [26]. A overview of our system is shown in Figure. 1.

3. EXPERIMENTS

3.1. Dataset and Training Setup

We train our model on a real spatial audio and visual recordings. The STARSS23[] dataset contains multichannel recordings of sound scenes in various rooms and environments, together with temporal and spatial annotations of prominent events belonging to a set of target classes. The original 360-degree video are spatially and temporally aligned with the microphone array recordings.

3.2. Experimental Results

We rigorously reproduced the challenge baseline framework to ensure fair comparison. As shown in Table. 1, our evaluation on the official STARSS23 development-test partition demonstrates the comparative performance between the proposed audio-visual system and the replicated baseline under standardized challenge metrics. Our results demonstrate overall superiority over the replicated baseline across most metrics, with the exception of a marginal deficit in onscreen accuracy (OSA).

4. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[2] L. Gao, Q. Mao, and M. Dong, "On local temporal embedding for semi-supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1687–1698, 2024.

[3] L. Gao, Q. Mao, M. Dong, Y. Jing, and R. Chinnam, "On learning disentangled representation for acoustic event detection," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2006–2014. [Online]. Available: https://doi.org/10.1145/3343031.3351086

[4] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[5] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.

[6] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[7] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 286–290.

[8] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, *et al.*, "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in neural information processing systems*, vol. 36, pp. 72 931–72 957, 2023.

[9] D. Berghi and P. J. Jackson, "Leveraging reverberation and visual depth cues for sound event localization and detection with distance estimation," *arXiv preprint arXiv:2410.22271*, 2024.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] A. Zhang and C. Li, "Scene text recognition based on multi-head attention fusion," *Radio Engineering*, vol. 54, no. 11, pp. 2576–2584, 2024.

[13] M. Lin, Z. Wang, W. Chen, J. Wu, and Z. Wang, "Ffstie: Video restoration with full-frequency spatio-temporal information enhancement," *IEEE Signal Processing Letters*, vol. 32, pp. 1–5, 2025.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the Conference on Association for Computational Linguistics*, 2019, pp. 6558–6569.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Self-attention with relative position representations," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 464–468.

[20] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal self-supervised learning from general videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 730–747.

[21] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[22] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal self-supervised learning from general videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 730–747.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[25] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," *Proceedings of the 30th International Conference on Machine Learning*, pp. 1058–1066, 2013.

[26] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 639–658.