Occupancy-Based Semantic Acoustic Imaging using NGCC-PHAT and Mel-Spectrogram Feature Fusion

Partha Pratim Deka
Computer Science and Engineering
Guwahati, India
parthapratimdeka140@gmail.com

***Abstract*—Semantic Acoustic Imaging (SAI) aims to reconstruct spatially-resolved acoustic representations from low-channel microphone recordings, enabling simultaneous sound event localization and semantic understanding. The DCASE 2026 SAI-SELD challenge introduces a particularly challenging setting where high-resolution acoustic maps must be inferred from four-channel spatial audio. In this work, we propose an audio-only framework that combines Neural Generalized Cross-Correlation with Phase Transform (NGCC-PHAT) features and Mel-spectrogram representations to jointly capture spatial and spectral characteristics of acoustic scenes. The NGCC-PHAT branch provides robust inter-channel localization cues, while the Mel-spectrogram branch captures event-specific spectral information. These complementary representations are fused and processed using a transformer-based prediction network to generate class-wise acoustic localization maps for the target sound events. During inference, localization peaks are converted into challenge-compliant acoustic map representations for evaluation. Experimental results on the STARSS23 development dataset demonstrate the feasibility of combining spatial correlation features and spectral representations for audio-only semantic acoustic imaging. The proposed framework demonstrates an occupancy-based formulation for semantic acoustic imaging from audio-only observations.**

I. Introduction

Sound Event Localization and Detection (SELD) aims to identify active sound events and estimate their spatial locations from multichannel audio recordings. Traditional SELD systems typically represent source locations using direction-of-arrival (DOA) estimates, resulting in sparse point-based spatial representations. The DCASE 2026 Semantic Acoustic Imaging for Sound Event Localization and Detection (SAI-SELD) task extends this paradigm by requiring systems to reconstruct dense semantic acoustic maps that jointly encode sound event categories, spatial locations, and acoustic energy distributions from low-channel spatial audio recordings. This formulation enables a richer representation of acoustic scenes and supports applications such as robotic perception, smart environments, and audiovisual scene understanding.

The challenge is particularly difficult because systems are required to infer high-resolution acoustic representations from only four-channel microphone recordings. Existing SAI systems often construct intermediate acoustic image representations through spatial upscaling networks before performing localization and segmentation. While effective, such approaches introduce additional computational complexity and may propagate errors from the acoustic image reconstruction stage into downstream localization models. While these representations provide useful localization information, they often emphasize either spatial structure or spectral content, limiting their ability to jointly model sound event semantics and source localization.

The official DCASE baseline [1] follows this general paradigm by first generating acoustic image representations and then performing instance-level localization and segmentation. Motivated by the complexity of this two-stage formulation, we investigate a direct occupancy prediction approach that operates entirely in feature space.

To address these challenges, we propose a spatial-spectral fusion framework that combines Neural Generalized Cross-Correlation with Phase Transform (NGCC-PHAT) [2] features and Mel-spectrogram representations for audio-only semantic acoustic imaging. NGCC-PHAT captures inter-channel time-delay information and spatial correlations that are critical for source localization, while Mel-spectrograms provide complementary spectral information for sound event discrimination. The two feature representations are fused and processed using a Transformer-encoder [3] based architecture capable of modeling both local acoustic patterns and long-range contextual dependencies.

Direct prediction of high-resolution acoustic maps introduces a highly sparse learning problem, as most spatial locations contain little or no acoustic energy. To reduce prediction complexity, the proposed framework predicts low-resolution class-wise occupancy maps that represent source presence and approximate spatial location. These occupancy predictions are subsequently converted into challenge-compliant acoustic map representations during inference. The main contributions of this work are summarized as follows:

- We propose a spatial-spectral feature fusion framework that combines NGCC-PHAT and Mel-spectrogram representations for audio-only semantic acoustic imaging.
- We introduce a Transformer-based acoustic imaging architecture for jointly modeling localization and semantic information from fused spatial and spectral features.
- We formulate semantic acoustic imaging as a low-resolution occupancy estimation problem, reducing the dimensionality of the prediction space and simplifying the learning objective compared to direct dense acoustic map regression.

II. Proposed Method

The proposed framework directly predicts semantic acoustic localization maps from four-channel spatial audio without constructing intermediate acoustic images. Figure 1 presents an overview of the proposed pipeline. First, spatial and spectral representations are extracted using NGCC-PHAT and Mel-spectrogram features, respectively. These complementary features are fused and processed by a Transformer-encoder based network to generate class-wise occupancy maps. Finally, the occupancy predictions are converted into challenge-compliant acoustic map representations through occupancy-based post-processing.

A. Spatial-Spectral Feature Extraction

Accurate semantic acoustic imaging requires both localization and semantic information. To capture these complementary characteristics, we employ two feature representations derived from the input four-channel spatial audio.

1) NGCC-PHAT Features: Spatial information is extracted using Neural Generalized Cross-Correlation with Phase Transform (NGCC-PHAT). For each microphone pair, channel signals are processed through a shared feature extraction backbone followed by generalized cross-correlation in the frequency domain. The resulting correlation maps encode inter-channel time-delay information, which is strongly related to source direction.

Unlike conventional GCC-PHAT features computed directly from waveform signals, NGCC-PHAT operates on learned feature representations, enabling more robust localization cues under challenging acoustic conditions. Correlation features from all microphone pairs are concatenated to form a unified spatial representation.

2) Mel-Spectrogram Features: While NGCC-PHAT provides localization information, it contains limited semantic content regarding sound event categories. To complement the spatial representation, Mel-spectrogram features are extracted from the multi-channel recording.

The Mel representation captures frequency-dependent acoustic characteristics that are highly informative for sound event recognition. In the proposed framework, channel signals are first aggregated and transformed into a Mel-frequency representation, producing a compact spectral description of the acoustic scene.

B. Feature Fusion

The NGCC-PHAT and Mel-spectrogram representations capture complementary aspects of the acoustic environment. The former emphasizes spatial localization cues, whereas the latter focuses on spectral and semantic characteristics.

To jointly exploit both information sources, the extracted feature representations are concatenated along the feature dimension to produce a unified spatial-spectral representation. This fused representation serves as the input to the subsequent prediction network.

NGCC-PHAT features produce a representation of size 384×47, while the Mel-spectrogram branch produces a representation of size 64×47. The two representations are concatenated along the feature dimension to obtain a fused representation $F \in \mathbb{R}^{448\times47}$.

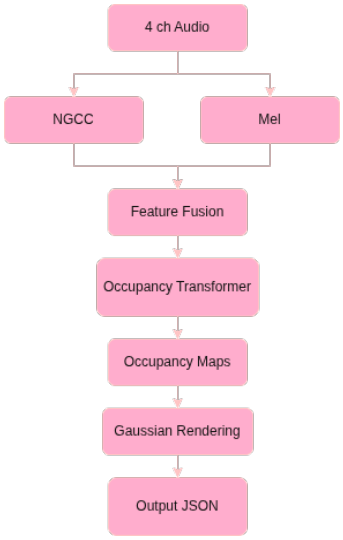

Fig. 1. Overview of the proposed SAI-SELD pipeline combining NGCC-PHAT and Mel-spectrogram features.

C. Transformer-encoder based Occupancy Prediction

The fused spatial-spectral representation is processed using a transformer encoder to model interactions between localization and semantic information. Compared to conventional convolutional architectures, transformers provide a larger receptive field and enable direct modeling of long-range dependencies across the feature sequence. This property is particularly beneficial for semantic acoustic imaging, where localization cues and spectral characteristics may be distributed across multiple temporal and frequency regions.

The prediction network first projects the fused feature representation into a shared latent embedding space through a series of convolutional layers. The resulting feature sequence is then processed by multiple transformer encoder blocks consisting of multi-head self-attention and feed-forward layers. Self-attention enables the model to dynamically capture relationships between spatial and spectral cues, allowing localization information from NGCC-PHAT features to be associated with semantic information encoded in the Mel-spectrogram representation.

The transformer output is subsequently passed through a fully connected prediction head that generates class-wise occupancy maps. Given C sound event classes, the network predicts an output tensor:

$$[O \in \mathbb{R}^{13\times30\times60}]$$

where each channel corresponds to a single sound event category and each spatial location represents the likelihood of the presence of that event at the corresponding acoustic position.

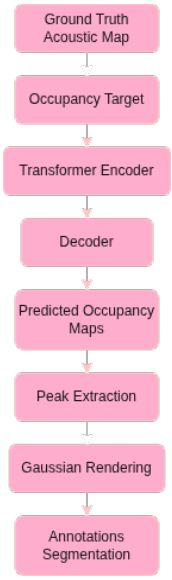

Fig. 2. Transformer-based occupancy prediction architecture.

Unlike dense acoustic image regression approaches, the proposed formulation focuses on estimating source occupancy and localization. This reduces the complexity of the learning problem by avoiding direct prediction of detailed acoustic energy distributions while preserving the spatial information required for downstream reconstruction and evaluation.

D. Acoustic Map Reconstruction

The objective of the SAI-SELD task is to generate semantic acoustic maps that describe both the spatial location and acoustic energy distribution of active sound sources. Directly learning these dense energy maps is challenging due to the large variability in source extent, intensity distribution, and acoustic propagation effects. Furthermore, the resulting acoustic maps often contain localized energy concentrations, making the prediction problem highly sparse.

To address this challenge, the proposed framework predicts class-wise occupancy maps instead of dense acoustic energy fields. Each channel of the output tensor represents a sound event category, while the activation values indicate the likelihood of source presence at each spatial location.

During inference, the predicted occupancy maps are converted into challenge-compliant acoustic annotations. For each sound event class, activated spatial regions are identified through thresholding of the occupancy map. The most confident activations are retained and transformed into sparse spatial point representations containing location and confidence information.

These sparse spatial predictions are subsequently converted into the acoustic map format required by the challenge evaluation framework. By predicting occupancy rather than directly regressing dense acoustic energy fields, the model focuses on estimating source presence and approximate location while delegating detailed acoustic map generation to the post-processing stage.

This formulation substantially reduces output dimensionality and alleviates the difficulty of learning highly sparse acoustic energy distributions directly from audio observations.

The proposed formulation offers two advantages. First, it significantly reduces the prediction complexity compared to dense acoustic map regression. Second, it aligns naturally with the evaluation procedure, which ultimately renders sparse source representations into continuous acoustic energy maps before metric computation. As a result, localization information can be estimated without requiring the network to explicitly learn detailed source energy shapes.

III. Experimental Setup

A. Dataset

Experiments were conducted on the STAIRS26 [4] development dataset, derived from STARSS23 [5] development dataset provided as part of the DCASE 2026 Task 3 Semantic Acoustic Imaging for Sound Event Localization and Detection (SAI-SELD) challenge. The dataset consists of spatial audio recordings captured using a 32-channel Eigenmike microphone array together with corresponding high-resolution acoustic map annotations. During evaluation, only four-channel spatial audio recordings are available, requiring models to reconstruct semantic acoustic maps from limited spatial information.

The benchmark contains 13 target sound event classes, including speech, music, footsteps, door events, and domestic sounds. Acoustic annotations are provided at a temporal resolution of 10 FPS and encode sound event category, spatial location, and acoustic intensity information.

B. Feature Extraction

The proposed framework utilizes complementary spatial and spectral representations extracted from the four-channel audio recordings.

For spatial representation learning, NGCC-PHAT features are extracted from microphone pair combinations. Each channel pair is processed using a shared neural feature extractor followed by generalized cross-correlation with phase transform (GCC-PHAT). The resulting correlation representations encode inter-channel delay information associated with source direction.

To capture semantic information, Mel-spectrogram features are computed from the audio signal using a sampling rate of 24 kHz. The Mel representation provides a compact description of the spectral content of the acoustic scene and complements the localization cues captured by NGCC-PHAT.

The spatial and spectral features are concatenated to form a unified spatial-spectral representation that serves as input to the prediction network.

C. Training Configuration

Audio recordings are segmented into fixed-length analysis windows centered around each annotation frame. The proposed transformer-encoder based network is trained to predict class-wise occupancy maps corresponding to the active sound events within each frame.

TABLE I
TRAINING HYPERPARAMETERS

Parameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	28
Epochs	20
Input Channels	448
Output Classes	13
Occupancy Height	30
Occupancy Width	60

TABLE II
DEVELOPMENT SET PERFORMANCE.

Metric	Value
Macro mAP	0.00019
Macro AP50	0.00014
Macro AP75	0.00000
Macro Pearson-r	0.0653
Micro mAP	0.00068
Micro Pearson-r	0.0397

The proposed model contains approximately 189.2 million trainable parameters.

The final network predicts occupancy maps of size $13 \times 30 \times 60$ corresponding to the 13 target sound event categories.

The model is optimized using the Adam optimizer with an initial learning rate of *1e-4*. Training is performed with a batch size of *28* for *20* epochs. Model selection is based on validation performance on the development set.

Table I summarizes the primary training hyperparameters.

D. Evaluation Metrics

Performance is evaluated using the official DCASE 2026 SAI-SELD evaluation framework. The primary ranking metric is Mask mean Average Precision (Mask mAP), which jointly evaluates sound event detection accuracy, localization quality, and acoustic map reconstruction performance.

Additionally, Pearson Correlation (r) is reported to measure energy field reconstruction fidelity, while Relative Distance Error (RDE) evaluates source distance estimation accuracy. Following the challenge protocol, predicted and reference annotations are rendered into continuous acoustic energy maps through spherical Gaussian kernels prior to metric computation.

IV. RESULTS AND DISCUSSION

A. Main Results

The proposed framework was evaluated on the STARSS23 development dataset using the official DCASE evaluation protocol.

The proposed approach demonstrates the ability to learn coarse spatial activation patterns directly from fused spatial-spectral audio representations without requiring intermediate acoustic image reconstruction. The results indicate that directly learning semantic acoustic representations from NGCC-PHAT and Mel-spectrogram features is sufficient for learning coarse localization patterns and sound-event-specific spatial activations.

B. Limitations

The proposed system was evaluated using the fused spatial-spectral representation only. Future work will investigate the individual contributions of NGCC-PHAT and Mel-spectrogram features through controlled ablation studies.

C. Discussion

The experimental results suggest that the proposed framework benefits from both the feature representation and the occupancy-based prediction strategy. Unlike image-based acoustic imaging approaches that first reconstruct intermediate acoustic representations, the proposed method directly learns sound event localization from compact spatial and spectral features.

The occupancy formulation further simplifies the learning objective by focusing on source presence and localization rather than dense acoustic energy reconstruction. This enables efficient prediction while remaining compatible with the challenge evaluation framework, which ultimately evaluates rendered acoustic energy maps.

Although the resulting localization accuracy remains limited, the experiments indicate that occupancy-based prediction is a viable formulation of semantic acoustic imaging. Future work will investigate improved occupancy supervision, higher-resolution spatial representations, and more effective localization-aware loss functions.

V. CONCLUSION

This paper presented an audio-only semantic acoustic imaging framework for the DCASE 2026 SAI-SELD challenge. The proposed approach combines NGCC-PHAT spatial features and Mel-spectrogram spectral representations to jointly capture localization and semantic information from four-channel spatial audio recordings.

A transformer-based prediction network was employed to generate class-wise occupancy maps, which were subsequently converted into challenge-compliant acoustic map representations through occupancy-based post-processing. By operating directly in feature space, the proposed framework avoids the need for intermediate acoustic image generation while preserving the spatial information required for semantic acoustic imaging.

Experimental results demonstrate the feasibility of occupancy-based semantic acoustic imaging using fused spatial and spectral representations. Future work will investigate improved occupancy estimation techniques, temporal modeling strategies, and more advanced feature fusion mechanisms for semantic acoustic imaging.

REFERENCES

[1] A. S. Roman, I. R. Roman, and J. P. Bello, "Latent acoustic mapping for direction of arrival estimation: A self-supervised approach," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2025, pp. 1–5.

[2] A. Berg, M. O'Connor, K. Åström, and M. Oskarsson, "Extending GCC-PHAT using Shift Equivariant Neural Networks," in *Proc. Interspeech 2022*, 2022, pp. 1791–1795.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf
[4] I. R. Roman, A. Politis, K. Shimada, H. Cheston, P. Sudarsanam, D. Díaz-Guerra, Y. Sun, T. Shibuya, S. Takahashi, and Y. Mitsufuji, "Stairs26: Sony-tau acoustic images of real-world scapes," Apr. 2026. [Online]. Available: https://doi.org/10.5281/zenodo.18171005
[5] 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 *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957.