AN EQUIRECTANGULAR ENERGY FIELD WITH PER-LOCATION CLASS COVERAGE FOR SEMANTIC ACOUSTIC IMAGING

Technical Report

Seunggyu Jeong[1,2]*, Seong-Eun Kim*[1,2]

[1] Medisensing, Seoul, Korea
[2] Seoul National University of Science and Technology, Seoul, Korea
wa3229433@gmail.com, sekim@seoultech.ac.kr

ABSTRACT

We describe our submission to DCASE 2026 Challenge Task 3, semantic acoustic imaging for sound event localization and detection on the audio-only track. For each frame the task asks for a set of acoustic-image instances, each a soft energy footprint on the equirectangular sphere together with a sound class, scored by a mask mean average precision (mask mAP). We factor the problem into a localization stage that predicts a dense energy field and a classification stage that labels each extracted instance. The localization stage uses a distilled acoustic-imaging front-end followed by PanoFormer, an equirectangular transformer whose circular padding removes the seam artefacts that a planar decoder produces at the azimuth wrap. Because the official MACRO average counts a class with no predictions as zero, the classification stage assigns a class to every instance from the direction of arrival of a pretrained spatial network, which covers all thirteen classes. A temporal-persistence re-ranking sets the detection confidences. The primary system reaches a development MACRO mask mAP of 0.0515, against 0.0003 for the official baseline. We submit four systems that differ in the classification stage.

***Index Terms*—** Acoustic imaging, sound event localization and detection, equirectangular transformer, spherical energy field, direction of arrival

1. INTRODUCTION

DCASE 2026 Task 3 [1] formulates semantic acoustic imaging for SELD. For every 0.1 s frame a system predicts a set of acoustic-image instances, where each instance is a soft energy footprint on the equirectangular sphere ($360^\circ \times 180^\circ$) together with a sound class. Predicted and reference footprints are matched by a soft spherical intersection over union under a $\sigma{=}6^\circ$ Gaussian render, with a 20° great-circle gate and no cross-class matches, and the score is averaged over the intersection-over-union thresholds $\{0.25, 0.5, 0.75\}$. The headline number is the MACRO mask mAP, the mean of the per-class average precisions. A class for which a system emits no prediction contributes zero to this mean rather than being skipped, so a system has to label all thirteen classes and not only the frequent ones.

The data are spatial scene recordings from the STARSS line [2, 3], with additional synthetic scenes generated by AudibleLight [4].

We work on the audio-only track and split the problem into localization, a dense energy field over the sphere from which instances are taken as connected components, and classification, a sound class for each instance. The official baseline builds the field with a ResNet/FPN decoder on top of an acoustic-imaging front-end. Two properties of the metric guide our design. First, every spurious connected component is a false positive, so the field has to be clean. Second, the MACRO average rewards covering every class. Section 2 describes the field, Section 3 the classification and confidence stages, Section 4 the four submitted systems, and Section 5 development results.

2. LOCALIZATION: AN EQUIRECTANGULAR ENERGY FIELD

Front-end. The localization branch reads a multi-band acoustic energy image produced by a distilled imaging front-end (UpLAM) built on latent acoustic mapping [5] that maps the four-channel first-order recording to a nine-band equirectangular image of size $9 \times 180 \times 360$, one band per frequency sub-range. The front-end is a $32 \rightarrow 4$ channel distillation of a conventional beamformer and is kept fixed. The raw image spans a very large dynamic range, so a per-channel instance normalization precedes the backbone to keep training stable.

Backbone. The energy field is a spherical signal whose left and right borders are adjacent in azimuth. A planar convolutional decoder treats that border as a discontinuity and emits wrap artefacts that become false-positive components. We replace the ResNet/FPN decoder with PanoFormer [6], an equirectangular transformer whose attention and patch handling include azimuth-circular padding. It regresses a single-channel dense field at 512×1024, which we down-sample to the 180×360 scoring grid. The target field places a spherical Gaussian at each annotated source, with a higher loss weight on the source pixels to balance foreground and background. The equirectangular field is markedly cleaner than the planar one and, in the submittable setting where every connected component is kept, raises the mask mAP from about 0.017 to 0.0394 at a fixed operating point.

Instance extraction. Instances are connected components of the field thresholded at 0.5. Each component is exported as up to 60 points carrying their energy, and its detection score starts from the peak energy, which we verified to be a better ranking statistic than sum, mean or area-weighted pooling. The threshold and the number of points per instance form the operating point; on the development set a threshold of 0.5 with 60 points is the best setting under the 20 MB per-file submission limit.

Table 1: Submitted systems and development MACRO mask mAP. All share the PanoFormer field and temporal-persistence confidence and differ in the class branch. "Strict" counts a class with no predictions as zero (the official average); "lenient" averages only over predicted classes.

System	Class branch	lenient	strict
1	frozen SSAST probe	0.0710	0.0156
2	PSELDNets coverage, 80px radius	0.0515	**0.0515**
3	CLAP present set + PSELDNets	0.0415	0.0415
4	PSELDNets coverage, 40px radius	0.0501	0.0501

3. CLASSIFICATION AND CONFIDENCE

Per-frame probe. A first class estimate comes from SSAST [7], a self-supervised audio spectrogram transformer pre-trained on AudioSet [8]. We keep the backbone frozen and train a thirteen-class linear head on its time-pooled embedding. Fine-tuning the backbone (full or low-rank) raised the training accuracy but lowered the test mask mAP, so we use the frozen probe.

Per-location coverage. On the real recordings the per-frame probe concentrates on a few frequent classes, which is adequate under a predicted-classes-only average but weak under the official MACRO, where the missing classes score zero. To label every class we use PSELDNets [9], an HTSAT-ACCDOA spatial network pre-trained on large-scale synthetic spatial audio and fine-tuned here with a thirteen-class ACCDOA head. For each frame it produces a per-class direction of arrival from the first-order recording, and each extracted instance takes the class of the nearest active direction within a per-class assignment radius. Because the four-channel direction estimate is coarse, a wide radius for the coverage classes recovers more of their true detections; on the development set a radius of 80 pixels for those classes, against 40 for the frequent ones, gives the best MACRO mask mAP. This stage labels all thirteen classes and is the primary classification stage.

Confidence. The detection confidences are set by a temporal-persistence re-ranking. Each detection is scored by its peak energy multiplied by the number of nearby frames within 20° that carry a co-located detection of the same class, which moves short-lived false positives down the ranking and improves the average precision. This re-ranking is applied to every submitted system.

4. SUBMITTED SYSTEMS

The four systems share the localization field and the temporal-persistence confidence, and differ in how the class of each instance is set. Table 1 lists them. System 2 is the primary system: the per-location PSELDNets coverage labels all thirteen classes and, with the wide assignment radius, gives the best MACRO mask mAP. System 1 uses the frozen SSAST probe, which is the strongest system under a predicted-classes-only average and provides a different class distribution. System 3 forms a per-frame present-class set with CLAP [10], an audio-language model that recovers transient classes well in isolation, and assigns each instance the nearest PSELDNets direction within that set. System 4 is the same coverage system at the narrower 40-pixel radius, which keeps only the more confident assignments. The set covers both averaging conventions and two assignment radii for the unseen evaluation set.

5. RESULTS

Table 1 reports development results in the submittable setting, where every connected component is kept, so the numbers reflect what the pipeline produces on unseen data. The localization upgrade from the ResNet/FPN field to the equirectangular PanoFormer field is the main gain. Under the official MACRO average the per-location PSELDNets coverage gives the primary system at a MACRO mask mAP of 0.0515, 172 times the official baseline of 0.0003, with a per-recording size below the 20 MB limit. Assigning every instance its true class on the same field would reach 0.0799, which bounds what the classification stage can add and leaves localization accuracy as the main remaining limit.

6. CONCLUSION

Our Task 3 system predicts a clean equirectangular energy field, extracts instances as connected components, labels every instance from a per-location direction of arrival so that all thirteen classes are covered, and sets the confidences by temporal persistence. The equirectangular field is the main localization gain, and the direction-of-arrival coverage is what the MACRO average rewards. We submit four systems that differ in the class branch, spanning the two averaging conventions and several class distributions for the unseen evaluation set.

7. REFERENCES

[1] DCASE Community, "DCASE 2026 challenge task 3: Semantic acoustic imaging for sound event localization and detection," https://dcase.community/challenge2026/, 2026.

[2] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129.

[3] 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*, vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957.

[4] H. Cheston, A. Stepien, J. Azcarreta, A. S. Roman, C. Chen, C. Bilen, and I. R. Roman, "AudibleLight: A controllable, end-to-end API for soundscape synthesis across ray-traced and real-world measured acoustics," in *DMRN+20: Digital Music Research Network One-Day Workshop 2025*, 2025.

[5] 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.

[6] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, "PanoFormer: Panorama transformer for indoor 360° depth estimation," in *European Conference on Computer Vision (ECCV)*, 2022.

[7] Y. Gong, C.-I. J. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[8] J. F. Gemmeke, D. P. W. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] J. Hu, Y. Cao, M. Wu, *et al.*, "PSELDNets: Pre-trained neural networks on a large-scale synthetic dataset for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2024.

[10] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.