

BEARINGS: SELF-SUPERVISED SOUNDFIELD EMBEDDINGS FROM FIRST-ORDER AMBISONICS

Goksenin Yuksel¹, Marcel van Gerven¹, Kiki van der Heijden^{1,2}

¹Dept. of Machine Learning and Neural Computing, Donders Institute,
Radboud University, Nijmegen, the Netherlands

²Mortimer B. Zuckerman Institute, Columbia University, New York, USA

ABSTRACT

Recently proposed self-supervised audio encoders learn powerful general-purpose representations of sound scenes, yet they are spatially blind. To supply the missing spatial representation of sound scenes, we introduce BEARINGS. BEARINGS is a self-supervised framework that learns soundfield embeddings from unlabeled first-order Ambisonics. We pre-train a masked auto-encoder paired with a decoder conditioned on frozen acoustic embeddings from an off-the-shelf single-channel audio encoder. Our results show that the resulting soundfield embeddings form a reusable stream that can be attached to frozen acoustic encoders with a lightweight trainable fusion head. On sound event localization and detection, concatenating our soundfield embeddings with acoustic representations provides the missing spatial information and enables joint detection and localization, raising the location-dependent F-score from below 4 to 50 on TAU-NIGENS 2021 and 39 on STARSS23. To our knowledge, BEARINGS is the first self-supervised soundfield encoder whose embeddings plug into frozen acoustic encoders without retraining either model.

Index Terms— Self-supervised learning, Spatial audio, Ambisonics, Sound event localization and detection

1. INTRODUCTION

Self-supervised general-purpose audio representation learning has seen remarkable progress: models such as ATST [1], BEATs [2], Dasheng [3], SPEAR [4] learn powerful general-purpose embeddings from large-scale unlabeled audio. Despite their acoustic fidelity, these models are inherently blind to spatial information: they encode only *what* sounds are present, not *where* those sounds originate.

Joint spatial and acoustic perception has traditionally been addressed through supervised sound event localization and detection (SELD) [5], where models such as SELDNet [6] map multichannel audio features to joint sound detection and direction-of-arrival estimates. More recently, pre-training has been extended to spatial audio: w2v-SELD [7] adapts contrastive waveform pre-training to Ambisonics for SELD; Spatial-AST—the spatial front-end of BAT [8]—is trained on synthetic binaural mixtures for joint localization, detection, and LLM-based spatial reasoning; and ELSA [9] aligns spatial audio with text representations for 3D descriptions of the sound scenes. Closest to our setting, GRAM-Ambisonics [10] extends self-supervised pre-training to Ambisonics, learning content and spatial structure jointly. These approaches entangle spatial and acoustic modeling within a single backbone [8, 10, 9]. The SELD approaches that were trained on real-life recordings are optimized for

a particular SELD setup [7, 6]. Concurrently, the emerging class of audio-LLMs [11, 12] has advanced high-level acoustic reasoning, yet they remain spatially constrained by their reliance on single-channel encoders, with only limited attempts at spatial grounding [8, 13]. While such spatial audio-LLMs can reason over static sound scenes, they do not expose spatial representations of the sound scene.

We introduce BEARINGS, a modular framework that learns reusable soundfield embeddings through spatial self-supervision. Once pre-trained, the soundfield encoder can be paired with different single-channel acoustic encoders while keeping both encoders frozen. This allows the downstream acoustic encoder to be replaced as stronger models become available, without repeating spatial pre-training.

We learn soundfield embeddings in a fully self-supervised manner from unlabelled in-the-wild clips crawled from 360° YouTube videos [14]. From the FOA signal, we extract a 7-channel spatial representation consisting of four log-mel spectrograms and three normalized active-intensity vectors. The soundfield encoder is trained as a masked auto-encoder whose decoder cross-attends to acoustic embeddings from a frozen single-channel encoder while reconstructing the direction field and diffuseness [15] (Fig. 1).

Downstream fusion experiments validate that BEARINGS¹ embeddings capture substantial spatial information. Fused with monaural acoustic content embeddings via lightweight heads, they raise the location-dependent F-score from below 4 to 41–50 on TAU-NIGENS 2021, and 34–39 on STARSS23 across three architecturally distinct acoustic backbones. With GRAM-Clean as the content encoder, the fused model surpasses the supervised SELDNet baseline [6] on every metric of both datasets except LE_{CD} on STARSS23, where the two tie at 22.0°, and surpasses the jointly trained GRAM-Ambisonics [10] on TAU-NIGENS 2021 while matching it on STARSS23. These results establish that soundfield representations can be learned without spatial annotations and reused across frozen acoustic encoders.

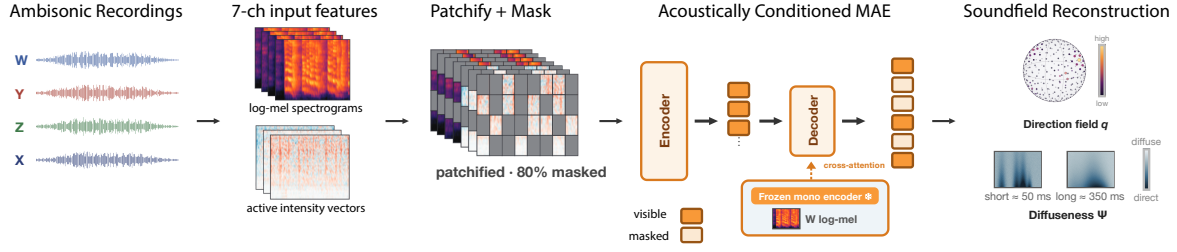
2. METHODOLOGY

2.1. Pre-Training Data

We pre-train the soundfield encoder on YT-AmbiGen [14], a dataset containing roughly 102 K five-second audio clips in first-order Ambisonics (FOA) format, which are sourced from 360° YouTube videos. The FOA signal consists of an omnidirectional channel $B_{00}(t)$ and three directional channels $B_{1,-1}(t)$, $B_{10}(t)$, $B_{11}(t)$. The clips capture real-world sound scenes with dynamic as well as

¹Code&Models: <https://github.com/labhamlet/Bearings>

A Self-Supervised Pre-Training



B Downstream Fusion (Joint Acoustic + Spatial Representations)

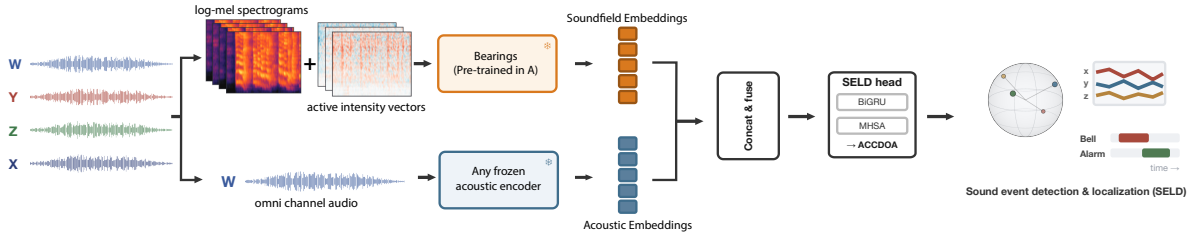


Fig. 1. BEARINGS overview. **(A)** Self-supervised pre-training: a masked ViT encodes seven-channel FOA features (four log-mel spectrograms, three active-intensity vectors), and a decoder that cross-attends to embeddings from a frozen mono encoder reconstructs direction distributions and two-scale diffuseness at the masked patches. **(B)** Downstream fusion: the frozen BEARINGS encoder supplies soundfield embeddings that are concatenated with acoustic embeddings from any frozen single-channel acoustic encoder.

static sound sources, exposing the model to naturalistic soundfields during pre-training.

2.2. Input Features

The soundfield encoder takes as input a 7-channel spatial time-frequency representation $\mathbf{F}(t, b) \in \mathbb{R}^{7 \times T \times B}$ extracted from the FOA signal, consisting of four log-mel spectrograms and three normalized active-intensity vectors. First, we normalize the FOA waveforms by B_{00} loudness. We then compute the log-mel spectrogram per channel (short-time Fourier transform, window size = 1024, hop length = 10 ms, 128 mel bands). Additionally, we compute the Cartesian active intensity at each time–frequency bin as

$$\mathbf{I}_a(t, f) = 2 \operatorname{Re} \left[\hat{B}_{00}^*(t, f) \cdot (\hat{B}_{1,-1}, \hat{B}_{10}, \hat{B}_{11})^\top \right], \quad (1)$$

with total energy $E(t, f) = |\hat{B}_{00}|^2 + |\hat{B}_{1,-1}|^2 + |\hat{B}_{10}|^2 + |\hat{B}_{11}|^2$. We aggregate intensity and energy to the mel scale separately and normalize per band, resulting in the active-intensity vector (AIV)

$$\hat{\mathbf{I}}_a(t, b) = \frac{\sum_f m_b(f) \mathbf{I}_a(t, f)}{\sum_f m_b(f) E(t, f) + \epsilon}. \quad (2)$$

2.3. Soundfield Encoder Architecture

We tokenize the input with a single non-overlapping 2D convolution that projects each time–frequency patch (shape $f \times t$ with $f = 16$ and $t = 8$) across all $C = 7$ channels jointly to dimension D . We add fixed 2D sine–cosine positional embeddings and pass the sequence through the soundfield encoder f_ψ , which is a shallow standard ViT backbone [16] (6 layers, width $D=384$, 10.9M parameters). The resulting patch embeddings $\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_P \in \mathbb{R}^D$ serve as frame-level soundfield embeddings for downstream tasks.

2.4. Acoustically-Conditioned Masked Reconstruction

We train f_ψ as a masked auto-encoder (MAE) [17]. The encoder operates on a small set of visible patches while the decoder reconstructs the spatial field at masked locations conditioned on acoustic embeddings extracted from a single-channel acoustic encoder.

Masking and encoding. We mask 80% of the P patches (tube masking across all seven channels at each masked location) and encode only the visible patches, yielding $\mathbf{z}_{\text{vis}} \in \mathbb{R}^{P_{\text{vis}} \times D}$. We use a masking strategy that masks (i) the full time axis of the input features (40%), (ii) random patches (20%), and (iii) right-censored future frames (40%).

Conditioned decoder. The encoder outputs and a shared learnable mask token at masked positions are fed into the decoder together with 2D sine–cosine positional embeddings. The decoder is a 4-layer ViT (width $D = 256$, 8 attention heads, 4.6 M parameters). Each decoder block applies (i) self-attention over the patch embeddings and (ii) cross-attention into a single-channel acoustic embedding sequence $\mathbf{C} = g_\theta(\text{Enc}_{\text{mono}}(M_{00}))$ obtained by passing the B_{00} log-mel spectrogram through a frozen GRAM-Clean encoder (the mono encoder of GRAM) [10].

Soundfield reconstruction. At each masked patch the decoder predicts two analytically derived spatial features.

(i) *Direction.* We predict a categorical distribution over a Fibonacci grid $\mathcal{G}$ of $G=256$. The target is an energy-weighted kernel density estimate of the local intensity direction: after smoothing the mel intensity over a small time–frequency tile (shared with the short diffuseness window below), each tile contributes a von Mises–Fisher bump centred on its smoothed direction $\mathbf{u}_\tau = \bar{\mathbf{I}}_\tau / \|\bar{\mathbf{I}}_\tau\|$, weighted by its smoothed intensity magnitude,

$$q_{p,g} \propto \sum_{\tau \in p} \|\bar{\mathbf{I}}_\tau\| \exp(\kappa (g^\top \mathbf{u}_\tau - 1)), \quad \kappa = G/2\pi. \quad (3)$$

The target is normalized so that $\sum_g q_{p,g} = 1$. We supervise it with a weighted cross-entropy over the masked set,

$$\mathcal{L}_q = \frac{\sum_{p \in \mathcal{M}} \omega_p (-\sum_g q_{p,g} \log \hat{q}_{p,g})}{\sum_{p \in \mathcal{M}} \omega_p}, \quad \omega_p = \frac{W_p}{W_p + \bar{W}}, \quad (4)$$

where $W_p = \sum_{\tau \in p} \|\bar{\mathbf{I}}_\tau\|$ is the patch’s accumulated directional energy and $\bar{W}$ the batch mean of W_p .

(ii) *Diffuseness (two time scales)*. Following the DirAC [15] formulation, diffuseness is one minus the ratio of the magnitude of the time-averaged active intensity to the time-averaged energy,

$$\Psi^{(s)}(t, b) = 1 - \frac{\|\langle \mathbf{I}_a^{\text{mel}} \rangle_{\tau_s}\|}{c \langle E^{\text{mel}} \rangle_{\tau_s} + \epsilon}, \quad (5)$$

computed from the *raw* mel intensity and energy, with $c = 1$ under SN3D normalization guaranteeing $\Psi \in [0, 1]$. We predict two scales, a short window $\tau_s \approx 50$ ms and a long window $\tau_l \approx 350$ ms. The loss is an ℓ_1 term, $\mathcal{L}_\Psi = \frac{1}{N} \sum_i |\hat{\Psi}_i - \Psi_i|$. The total objective is

$$\mathcal{L} = \mathcal{L}_q + \mathcal{L}_\Psi. \quad (6)$$

2.5. Downstream SELD Evaluation

We attach the frozen soundfield embeddings to frozen, state-of-the-art monaural acoustic encoders and train a lightweight head to perform SELD, a task that inherently requires joint reasoning over acoustic content and spatial directionality [5]. Here, the fusion combines acoustic embeddings with the spatial information captured by BEARINGS. We evaluate in two complementary settings: scenes generated with measured room impulse responses (TAU-NIGENS 2021 [18]) and real recordings captured with an Ambisonic microphone array (STARSS23 [19]). For STARSS23, we use the synthetic + real dataset as in the DCASE 2023 Challenge 3 baseline.

To test versatility, we pair the frozen soundfield encoder with three architecturally distinct frozen backbones, all of which process only the B_{00} channel: GRAM-Clean [10] (time–frequency patch tokens), Dasheng-Base [3] (time-only tokens), and SPEAR [4] (waveform). Tokens with an explicit frequency axis (BEARINGS, GRAM-Clean) are first pooled over frequency using learned frequency attention. All sequences are then temporally resampled to a 100 ms label grid by average pooling, projected to $d_{\text{proj}} = 256$ by separate linear layers, concatenated, and passed through a linear fusion layer. The fused sequence feeds a SELD head with the temporal architecture of the SELDNet [6] baseline (bidirectional GRU followed by multi-head self-attention), whose final linear layer produces ACC-DOA predictions per 100 ms frame.

Baselines. We compare the performance of BEARINGS against two baselines: (i) a fully supervised SELD reference trained from scratch on the 7-channel FOA input (SELNet, results taken from DCASE baselines [18, 19]), (ii) a self-supervised model pre-trained on 7-channel Ambisonics audio (log mel + IVs), evaluated with its encoder frozen under the same downstream protocol and SELD head (GRAM-Ambisonics [10]). Importantly, we compare against the DCASE baseline and GRAM-Ambisonics rather than challenge winning systems, which use ensembling, extensive data augmentation, external synthetic training data, and task-specific architectures trained end-to-end. Our claim concerns the transferability of frozen soundfield embeddings under a fixed lightweight head (1.3–1.9 M trainable parameters), not absolute SELD performance.

Ablation experiments. To isolate the benefit of the soundfield embeddings produced by BEARINGS on SELD performance, we compare performance to a single-channel, acoustic pipeline in which

the acoustic embeddings produced by the acoustic encoder are fed directly into the SELD head. To assess the impact of spatial pre-training, we replace the pre-trained soundfield encoder with a small trainable convolutional front-end (two Conv2D layers) that maps the 7-channel FOA input to spatial embeddings, which are fused with the acoustic embeddings.

Furthermore, to measure what spatial information is encoded by the embeddings produced by BEARINGS independently of any SELD head, we test sound localization performance and reverberation estimation (RT_{60}) in simulated spatial scenes. To this end, we render single-source FOA scenes in Matterport3D [20] houses with known source azimuth/elevation and reverberation time. We obtain clip-level soundfield embeddings by mean-pooling the BEARINGS embeddings. Each quantity is read out with a k -nearest-neighbour regressor fit on training houses and evaluated on held-out houses. As a reference, we apply the same readout to hand-crafted spatial statistics computed directly from the input features (mean and standard deviation of the active-intensity vectors, short- and long-window diffuseness averages and their difference, and B_{00} log-mel statistics).

Hand-crafted spatial features. As a parameter-free baseline, we assess performance when hand-crafted spatial statistics are fused to the single-channel, acoustic embeddings instead of the soundfield embeddings produced by BEARINGS. Spatial features consist of the framewise active-intensity vectors of Eq. (2), their 50 ms- and 350 ms-smoothed versions, and the magnitudes of the two smoothed vectors. For DOA and reverberation estimation in Matterport3D scenes, each clip is represented by the mean and standard deviation of its AIVs, average short- and long-window diffuseness and their difference, and statistics of the B_{00} log-mel spectrogram.

Metrics. We report the location-dependent error rate and F-score computed with a 20° spatial threshold (ER_{20° , F_{20°) together with the class-dependent localization error LE_{CD} (angular error in degrees; a class with no detection is assigned 180°) and localization recall LR_{CD} [18]. The 95% confidence intervals (Table 1) are jack-knife intervals computed over the test set. For localisation and reverberation estimation on the simulated spatial scenes (Matterport3D probe), we report median angular error and Pearson r for RT_{60} .

2.6. Implementation details

During pre-training, we randomly draw two 2 s crops from each 5 s clip. We pre-train models for 50 k steps with AdamW ($\beta_1=0.9$, $\beta_2=0.98$, peak learning rate 2×10^{-4} , weight decay 0.01, batch size 256, cosine decay after 10k warm-up steps). For the downstream evaluation, we follow the official DCASE baseline protocol of each dataset [18, 19]. Only the fusion layers and SELD head are trained (1.9 M trainable parameters with GRAM-Ambisonics, 1.3 M with SPEAR and Dasheng), using Adam (learning rate 0.001) with an MSE loss.

3. RESULTS

3.1. Downstream SELD Performance

Table 1 shows that single-channel acoustic encoders retain detection ability (LR_{CD} of 35–43 on TAU-NIGENS 2021) but do not localize them: location-dependent F-scores do not exceed 3.2, and angular errors are large. Adding BEARINGS embeddings to the single-channel encoders yields large gains for each encoder on both datasets without updating either encoder. With GRAM-Clean, the fused model surpasses SELDNet on every metric of both datasets except a tie on STARSS23 LE_{CD} (22.0°), and on TAU-NIGENS

Table 1. Downstream SELD performance. Bold marks the best value per column.

Backbone	Method	TAU-NIGENS 2021				STARSS23			
		ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑
	SELDNet (7-ch FOA + IVs) [18, 19]	0.69	33.9	24.1	43.9	0.57	29.9	22.0	47.7
	GRAM-Ambisonics (7-ch FOA + IVs) [10]	0.62	41.0	23.3	66.3	0.56	37.4	22.8	58.8
GRAM-Clean	Single-channel, acoustic only	0.83 [0.81,0.84]	2.8 [2.3,3.3]	102.1 [99.3,104.9]	34.7 [33.0,36.2]	0.83 [0.81,0.86]	2.5 [1.6,3.4]	147.8 [146.2,149.3]	16.6 [15.0,18.2]
	+ Bearings (ours)	0.55 [0.53,0.56]	50.4 [48.2,52.7]	19.3 [18.3,20.3]	66.8 [65.0,68.5]	0.51 [0.46,0.55]	38.9 [34.6,43.8]	22.0 [18.3,25.7]	55.5 [49.3,61.4]
Dasheng-Base	Single-channel, acoustic only	0.83 [0.82,0.85]	3.2 [2.6,3.9]	87.1 [84.1,90.3]	43.2 [41.5,45.0]	0.82 [0.80,0.86]	2.9 [2.3,3.5]	146.1 [144.6,147.5]	18.9 [17.1,20.7]
	+ Bearings (ours)	0.56 [0.55,0.58]	47.7 [45.7,50.1]	20.0 [19.0,21.0]	65.0 [62.9,67.2]	0.54 [0.50,0.58]	34.3 [30.2,38.7]	24.9 [21.5,29.3]	53.8 [48.7,59.1]
SPEAR-Base	Single-channel, acoustic only	0.89 [0.87,0.90]	3.0 [2.5,3.5]	86.4 [83.3,89.8]	36.5 [34.8,38.4]	0.86 [0.84,0.89]	2.4 [1.5,3.3]	145.9 [142.1,149.9]	11.3 [9.9,12.8]
	+ Bearings (ours)	0.61 [0.59,0.64]	40.9 [38.3,43.5]	22.2 [20.9,23.4]	61.5 [59.1,63.8]	0.56 [0.53,0.60]	34.9 [30.5,39.0]	21.2 [18.0,24.3]	49.0 [43.7,53.6]

Table 2. Label efficiency on TAU-NIGENS 2021.

Training data	Embeddings	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑
25%	Bearings (ours)	0.67	30.6	24.4	49.4
	Learned-In Domain	0.74	21.7	40.5	45.2
50%	Bearings (ours)	0.57	46.6	20.5	64.2
	Learned-In Domain	0.66	31.7	24.8	56.5
100%	Bearings (ours)	0.55	50.4	19.3	66.8
	Learned-In Domain	0.60	39.0	22.6	64.3

2021 surpasses GRAM-Ambisonics on every metric. On STARSS23 it improves ER_{20°} (0.51 vs. 0.56), while the intervals for F_{20°} and LE_{CD} contain GRAM-Ambisonics’s values and LR_{CD} is lower (55.5 vs. 58.8). Dasheng and SPEAR follow the same pattern with smaller margins: both surpass SELDNet on every TAU-NIGENS 2021 metric, with Dasheng within 2.7 F-score points of the GRAM-Clean pairing.

3.2. Ablation Studies

Label Efficiency. Table 2 shows that our soundfield embeddings consistently outperform the learned embeddings across all data fractions and metrics. Notably, with only 25% of the training data, our method achieves a relative improvement of 9.5% in ER_{20°} (0.67 vs. 0.74) and 41.0% in F_{20°} (30.6 vs. 21.7) over the baseline, demonstrating superior data efficiency. The advantage persists with 50% and 100% data, where our approach maintains higher localization accuracy and recall, as indicated by the improved *LE_{CD}* and *LR_{CD}* scores.

Components. Table 3 (row 2) shows that a randomly initialized, frozen soundfield encoder substantially underperforms on every metric, confirming that the fusion head reads out structure acquired during pre-training. Acoustic conditioning contributes consistently across metrics (row 3). Removing the direction loss collapses localization (DOA 49.2°, row 5) but slightly improves RT₆₀ correlation ($r=0.55$), suggesting the two objectives trade off. The diffuseness loss adds a modest SELD gain (F_{20°} 48.3 → 50.4, row 4). Finally, adding hand-crafted spatial features instead of the soundfield embeddings produced by BEARINGS results in similar performance on localization recall, but the BEARINGS soundfield embeddings remain ahead on every other metric (bottom row).

Impact of self-supervised pre-training. Table 4 varies the soundfield encoder’s initialization and downstream training setting while keeping the acoustic content encoder frozen. The pre-trained, frozen BEARINGS encoder outperforms its counterpart trained from

Table 3. Ablation studies (TAU-NIGENS 2021, Matterport3D).

Configuration	SELD				Probe	
	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	DOA° ↓	RT ₆₀ r ↑
Full model	0.55	50.4	19.3	66.8	16.1	0.53
Random init	0.70	25.8	28.2	55.4	30.4	0.26
No conditioning	0.57	46.4	20.6	65.2	17.5	0.50
w/o $\mathcal{L}_\Psi$	0.57	48.3	19.8	65.9	16.1	0.52
w/o $\mathcal{L}_q$	0.68	29.5	27.7	63.7	49.2	0.55
Hand-crafted	0.61	43.2	22.9	66.6	23.5	0.36

Table 4. Impact of spatial pre-training (TAU-NIGENS 2021).

Encoder initialization, downstream	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑
Single-channel, acoustic only	0.83	2.8	102.1	34.7
Bearings: random, frozen	0.70	25.8	28.2	55.4
Bearings: random, trained end-to-end	0.65	35.7	23.7	63.6
Bearings: Pre-trained, frozen	0.55	50.4	19.3	66.8
Bearings: Pre-trained, fine-tuned	0.54	51.2	18.1	65.5

scratch, increasing F_{20°} from 35.7 to 50.4 and reducing LE_{CD} from 23.7° to 19.3°. Under the tested training recipe, hand-crafted spatial features also outperform the encoder trained from scratch (Table 3). Fine-tuning the pre-trained encoder provides modest further improvements in F_{20°} (50.4 → 51.2) and LE_{CD} (19.3° → 18.1°), with a small decrease in LR_{CD} (66.8 → 65.5). These results demonstrate the benefit of spatial pre-training for downstream SELD and support reusing the learned soundfield embeddings with frozen encoders.

4. CONCLUSION

We introduced BEARINGS, a self-supervised framework that learns soundfield embeddings from FOA recordings by reconstructing masked direction fields and diffuseness. Attached to three architecturally distinct frozen single-channel encoders through a lightweight fusion head, the embeddings enable sound event localization and detection without retraining either model. These results show that spatial structure can be learned without spatial annotations, with the resulting encoder reused across different frozen acoustic backbones. Future work includes scaling the training corpus and integrating the soundfield embeddings into spatial audio LLMs.

5. ACKNOWLEDGMENTS

This project received funding from the NWO Talent Program (VI.Veni.202.184; KH). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-14624.

6. REFERENCES

- [1] Xian Li, Nian Shao, and Xiaofei Li, “Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1336–1351, 2024.
- [2] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei, “BEATs: Audio pre-training with acoustic tokenizers,” in *Proceedings of the 40th International Conference on Machine Learning*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, Eds. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193, PMLR.
- [3] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang, “Scaling up masked audio encoder learning for general audio classification,” in *Interspeech 2024*, 2024.
- [4] Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland, “SPEAR: A unified SSL framework for learning speech and audio representations,” in *Forty-third International Conference on Machine Learning*, 2026.
- [5] Jun-Wei Yeow, Ee-Leng Tan, Santi Peksi, and Woon-Seng Gan, “Environmental acoustic intelligence through sound event localization and detection: a review,” *npj Acoustics*, vol. 1, no. 1, pp. 31, Dec 2025.
- [6] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas 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, 2019.
- [7] Orlem Lima Dos Santos, Karen Rosero, Bruno Masiero, and Roberto de Alencar Lotufo, “w2v-SELD: A sound event localization and detection framework for self-supervised spatial audio pre-training,” *IEEE Access*, vol. 12, pp. 181553–181569, 2024.
- [8] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath, “BAT: learning to reason about spatial sounds with large language models,” in *Proceedings of the 41st International Conference on Machine Learning*. 2024, ICML’24, JMLR.org.
- [9] Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia, “Learning spatially-aware language and audio embeddings,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds. 2024, vol. 37, pp. 33505–33537, Curran Associates, Inc.
- [10] Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden, “GRAM: Spatial general-purpose audio representation models for real-world applications,” arXiv preprint arXiv:2506.00934, 2026.
- [11] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” arXiv preprint arXiv:2311.07919, 2023.
- [12] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang, “SALMONN: Towards generic hearing abilities for large language models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [13] Xilin Jiang, Hannes Gamper, and Sebastian Braun, “Sci-Phi: A large language model spatial audio descriptor,” *IEEE Open Journal of Signal Processing*, vol. 7, pp. 276–284, 2026.
- [14] Jaeyeon Kim, Heeseung Yun, and Gunhee Kim, “Towards scene-aware video-to-spatial audio generation,” *Int. J. Comput. Vision*, vol. 134, no. 4, Mar. 2026.
- [15] Ville Pulkki, “Spatial Sound Reproduction with Directional Audio Coding,” *Journal of the Audio Engineering Society*, vol. 55, pp. Paper 503–516, June 2007.
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.
- [17] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer, “Masked autoencoders that listen,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022, NIPS ’22, Curran Associates Inc.
- [18] Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen, “A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection,” in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 125–129.
- [19] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji, “STARSS23: an audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023, NIPS ’23, Curran Associates Inc.
- [20] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang, “Matterport3D: Learning from rgb-d data in indoor environments,” *International Conference on 3D Vision (3DV)*, 2017.